

# Toward a Visual Concept Vocabulary for GAN Latent Space

Sarah Schwettmann<sup>1</sup>, Evan Hernandez<sup>2</sup>, David Bau<sup>2</sup>, Samuel Klein<sup>3</sup>, Jacob Andreas<sup>2</sup>, Antonio Torralba<sup>2</sup>

<sup>1</sup>MIT BCS, <sup>2</sup>MIT CSAIL, <sup>3</sup>MIT KFG

{schwett, dez, davidbau, sjklein, jda, torralba}@mit.edu

## Abstract

A large body of recent work has identified transformations in the latent spaces of generative adversarial networks (GANs) that consistently and interpretably transform generated images. But existing techniques for identifying these transformations rely on either a fixed vocabulary of pre-specified visual concepts, or on unsupervised disentanglement techniques whose alignment with human judgments about perceptual salience is unknown. This paper introduces a new method for building open-ended vocabularies of primitive visual concepts represented in a GAN’s latent space. Our approach is built from three components: (1) automatic identification of perceptually salient directions based on their layer selectivity; (2) human annotation of these directions with free-form, compositional natural language descriptions; and (3) decomposition of these annotations into a visual concept vocabulary, consisting of distilled directions labeled with single words. Experiments show that concepts learned with our approach are reliable and composable—generalizing across classes, contexts, and observers, and enabling fine-grained manipulation of image style and content.

## 1. Introduction

GANs [8] map latent vectors  $\mathbf{z}$  to images  $\mathbf{x}$ . Past work has found that directions in this latent space can encode specific aspects of image semantics: StyleGAN trained on bedrooms, for example, contains a direction such that moving most  $\mathbf{z}$  in that direction causes indoor lighting to appear in the associated image [24]. However, current methods for identifying these directions are *ad hoc*, capturing only a limited set of human-salient dimensions of variation. In this paper, we describe how to construct more expressive and diverse sets of meaningful image transformations—a visual concept vocabulary—by decomposing freeform language descriptions of GAN transformations.

Consider trying to find a direction that makes an outdoor market more *festive* (Figure 1). The GAN latent space is too large to make random search feasible, while supervised

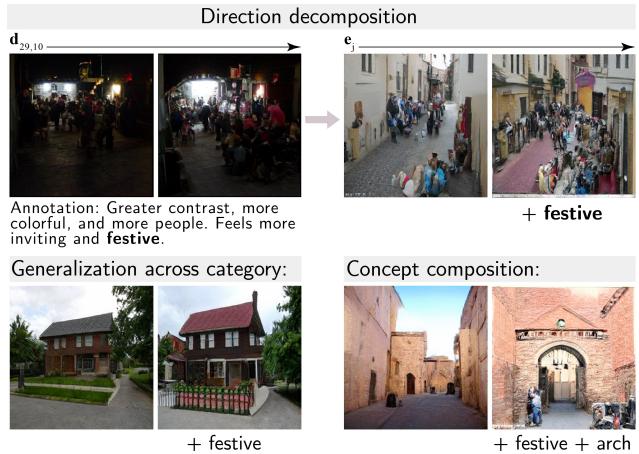


Figure 1: Building a visual concept vocabulary. First, we generate directions that preserve most of the structure and content in the image. Then we use human annotations to decompose them into directions that correspond to a single salient concept. Finally, we show the decomposed directions generalize across starting representations and input classes, and can be composed to construct compound directions.

approaches cannot verify if the desired direction is present [11, 7, 24, 19]. Unsupervised approaches [10, 15, 20, 21] may not discover a *festive* direction, since the model’s principal components do not necessarily capture changes that are most visually salient to humans.

To improve our understanding of the kinds of interpretable semantic transformations encoded in GAN latent space, we propose a new approach for building an open-ended glossary of primitive, perceptually salient directions from the bottom up. Our approach is built from three components:

1. A new procedure for generating perceptually salient directions based on **layer selectivity**. The resulting directions make meaningful local changes to a scene but are still non-atomic.
2. A data collection paradigm in which human annotators directly **label directions** with their semantics, which

are complex and compose multiple concepts to describe visual transformations.

3. A new **bag-of-directions model** which automatically decomposes these annotations into a glossary of “primitive” visual transformations associated with single words.

Because our method covers the breadth of the GAN latent space, it enables reliable image editing with a relatively open-ended vocabulary. We also show how our vocabulary supports generalization to novel compositions and transfer across classes. Code, data, and additional information are available at [visualvocab.csail.mit.edu](http://visualvocab.csail.mit.edu).

## 2. Related work

Our approach is inspired by recent success in discovering latent vectors that capture individual dimensions of semantic variation in images [11, 10, 7, 24]. To the best of our knowledge, ours is the first attempt to systematically catalog the set of human-interpretable concepts represented inside a generator’s latent space.

**Interpreting GANs.** GANs excel at capturing the rich visual structure of images—raising the question of what internal representations they leverage to do so, and the extent to which these representations overlap with dimensions of variation that humans recognize and find meaningful in visual scenes. Early work [17] on GANs discovered latent vectors that encode semantically meaningful representations at different levels of abstraction. A subsequent approach [2] to the interpretation problem focused on individual units, and used a pretrained segmentation network [23] to identify sets of units in intermediate layers whose feature maps closely match the semantic segmentation of a particular object class. Related work identified concepts *not* learned by GANs [3] by comparing the distribution of segmented objects in generated images with the target distribution in the training set. These approaches are constrained in the sets of concepts they could possibly identify, which are limited to the object classes represented inside the segmentation model. In addition to objects, GANs have also been shown to contain internal representations that determine spatial layout [14, 27, 1], and other higher-order scene attributes, including memorability and emotional valence [7]. While these approaches have made it possible to control specific aspects of image output, looking for a predetermined set of concepts limits what can be learned about what a GAN is able to represent. Our approach to interpretation aims to be more data-driven: by building *shared vocabularies*, represented by GANs and salient to humans, from the ground up.

**Supervised direction search.** If concepts to search for are known, and attribute annotations are available, vector directions in latent space can be discovered using supervised classifiers [12, 19]. When attribute annotations are not available, image classifiers can be used [24], or a separate model can be trained [11, 16, 6]. However the former is limited to concepts captured by the classifier, and the latter is limited to simple predetermined visual concepts, such as camera angle. Our method does not assume the concepts to search for are known ahead of time.

**Unsupervised direction search.** Other recent approaches use unsupervised methods for discovering interpretable dimensions in GAN latent space and feature space [10, 21, 15, 22]. These methods make use of the known disentanglement of many GAN representations [2]. One such method—GANSpace—discovers latent directions for image manipulation by identifying principal components of feature tensors on the early layers of GANs, and transferring the basis to latent space by linear regression [10]. However, the visual content of most of these transformations is unknown, as only a handful of examples have been labeled by the authors after the fact. Furthermore, this direction-generation procedure is limited to finding disentangled principal components of the model’s representation, while many other directions salient to humans may lie outside this set.

Where related work applies ad-hoc labels to directions discovered with such unsupervised methods, we introduce a bottom-up method for discovering directions associated with concepts, in the case when the set of concepts to search for is not known *a priori*. A primary contribution of our method is that it does not require visual concepts to be perfectly disentangled *before* labeling.

## 3. Projecting visual concepts into latent space

Our goal is to distill dimensions of variation in GAN latent space  $\mathcal{Z}$  that capture primitive visual transformations in image space. We begin by generating a set of test directions in multiple image classes, where transformation along those directions is constrained to be minimal in a subset of the layers’ feature representations, but potentially semantically complex (Section 3.1). Next, we synthesize sequences of images transformed along each test direction in  $\mathcal{Z}$  latent space, and ask human annotators to describe the corresponding visual changes (Section 3.2). Because these directions are generated without preselecting for particular concepts, they act as a screen upon which viewers project the gradients of perceptual change they find most salient. We use the prevalence of repeated terms and their association with different transformations to infer a set of visual concepts represented in the latent space, and associated directions that change the perceived presence of each concept (Section 3.3).

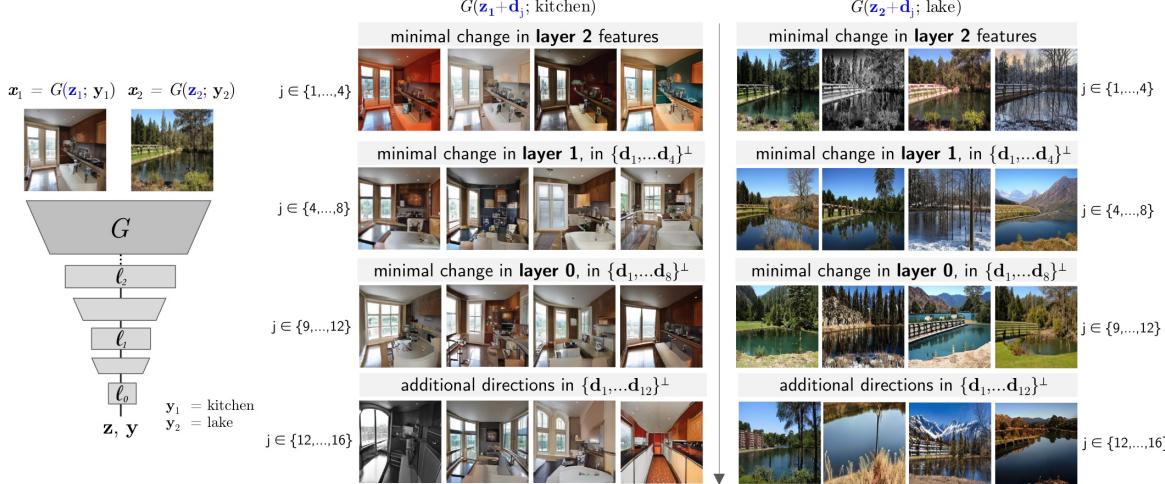


Figure 2: Examples of layer-selective directions. Directions are generated by minimizing the change in a layer with respect to the direction, subject to a norm constraint. Our procedure selects a set of  $n$  LSDs for each layer, one layer at a time, orthogonal to those already selected.

Experiments in the remainder of this paper use the BigGAN architecture [4], a class-conditional model pretrained on the Places dataset [26], which includes visual scenes from 365 unique classes. However, our approach is relatively model-agnostic. We show generalization to BigGAN trained on the ImageNet dataset [5] in the supplement.

### 3.1. Selecting directions for annotation

A generator  $G$  maps latent code  $\mathbf{z}$  and class vector  $\mathbf{y}$  into image space, synthesizing  $\mathbf{x} = G(\mathbf{z}; \mathbf{y})$ . The image  $\mathbf{x}$  can be manipulated along a visual dimension by transforming the vector  $\mathbf{z}$  along the corresponding direction  $\mathbf{d}$  in the latent space:  $\mathbf{x}^* = G(\mathbf{z} + \mathbf{d}; \mathbf{y})$ . This correspondence between directions in visual and latent space lies at the heart of the problem we wish to solve. For a given model, we want to learn embeddings in the latent space  $\mathcal{Z}$  of transformations that are salient to human observers in *visual space*. However, we cannot begin by defining an objective where  $\mathbf{d}$  is optimized to produce a discriminable transformation in  $\mathbf{x}$ , such as in [11, 7, 18, 24], because we wish to avoid pre-committing to a fixed vocabulary of visual concepts.

**Layer-selective directions (LSDs).** To generate directions for annotation, we sample the space of salient perceptual transformations for different  $\mathbf{z}$ . Our goal is to collect a direction annotation dataset that is both *diverse* and *specific*—capturing a broad set of concepts, where the same concept is reliably associated with a particular direction across images and observers. Thus for a given  $\mathbf{z}$ , we seek directions that make minimal, meaningful perceptual changes at different levels of abstraction.

Randomly sampled directions tend to alter many visual features, at many levels of resolution, all at once. To con-

strain a direction  $\mathbf{d}$  (of fixed magnitude) to make a smaller number of specific, recognizable changes in the image output  $G(\mathbf{z} + \mathbf{d}; \mathbf{y})$ , we can search for a  $\mathbf{d}$  that minimizes change in the feature representation of an intermediate layer of  $G$ . Denote by  $G_\ell$  the first  $\ell$  layers of  $G$ . Then the feature map for that layer is computed as follows:

$$g_\ell = G_\ell(\mathbf{z}, \mathbf{y}) \quad (1)$$

Let  $g_\ell^*$  be the output of layer  $\ell$  when we add  $\mathbf{d}$  to  $\mathbf{z}$ :

$$g_\ell^* = G_\ell(\mathbf{z} + \mathbf{d}, \mathbf{y}) \quad (2)$$

We constrain change in a layer’s representation by defining a layer regularizer that minimizes  $\|g_\ell^* - g_\ell\|^2$  for some layer  $\ell$ . To generate a diverse set of directions  $\mathbf{d}_{j,\ell}$  that meet this objective for layer  $\ell$ , we sample random vectors  $\mathbf{d}$  and then apply gradient descent to each sample to optimize the latent direction  $\mathbf{d}_{j,\ell}$  to minimize change in  $g_\ell^*$ , where  $\mathbf{d}_{j,\ell}$  is constrained to have unit norm. We call a direction optimized in this way a *layer-selective direction*.

Different layers of  $G$  control features in the image output at different levels of resolution, with later layers controlling more fine-grained features [2, 24]. Therefore, to construct a set of LSDs that encompasses diverse image transformations, when sampling vectors  $\mathbf{d}_{j,\ell}$  at layer  $\ell$  we add the further constraint that the samples be orthogonal to LSDs for other layers. Formally, our objective becomes:

$$\mathbf{d}_{j,\ell} = \arg \min_{\mathbf{d} \in U_\ell} \|g_\ell^* - g_\ell\|^2 \quad (3)$$

where  $U_\ell = \{\mathbf{d} \text{ such that } \|\mathbf{d}\| = 1 \text{ and}$

$$\mathbf{d} \perp \mathbf{d}_{j',\ell'} \text{ for all } j' \text{ and } \ell' > \ell\} \quad (4)$$

We begin by sampling  $n$  LSDs at the last layer  $\ell$  and then proceed to find orthogonal directions selective for earlier

layers. This procedure is analogous to Gram-Schmidt orthogonalization and picks directions that lie along mutually orthogonal subspaces of  $\mathcal{Z}$ , with transformations in each subspace corresponding to image changes at different levels of abstraction. Finally, we generate a set of  $n$  additional directions that are orthogonal to all LSDs, to capture types of image transformation that were excluded by the layer-selective process. Examples of directions generated using this method are visualized in Figure 2.

### 3.2. Collecting direction annotations

We apply the method described in Section 3.1 to 64 randomly selected  $\mathbf{z}$  to generate 20 layer-selective directions  $\mathbf{d}_j$  per  $\mathbf{z}$ , for a total of 1280 directions. For each  $\mathbf{z}_i$ , the image  $G(\mathbf{z}_i)$  is transformed along each direction  $j$  by passing a modified  $\mathbf{z}_i$  through the generator:  $G(\mathbf{z} + \alpha\mathbf{d}_{i,j})$ . The transformation is visualized in an image pair:  $[G(\mathbf{z}_i), G(\mathbf{z}_i + \alpha\mathbf{d}_{i,j})]$ , where  $\mathbf{d}$  has unit norm. To create images for annotation, we set the scaling term  $\alpha = 6$ .

For each direction, we synthesize images in four classes within BigGAN-Places: `cottage`, `kitchen`, `lake`, and `medina` (outdoor marketplace). These represent familiar visual scenes that balance indoor and outdoor, natural and built environments. Direction annotations are collected using Amazon Mechanical Turk (AMT). Participants see a single image pair  $[G(\mathbf{z}), G(\mathbf{z} + \alpha\mathbf{d}_j)]$  and are asked to describe the main visual changes in composition and style between the two images, for a total of 5,120 annotations. Figure 3 shows example images sequences and annotations. We provide further details on the AMT setup in Section S.1.

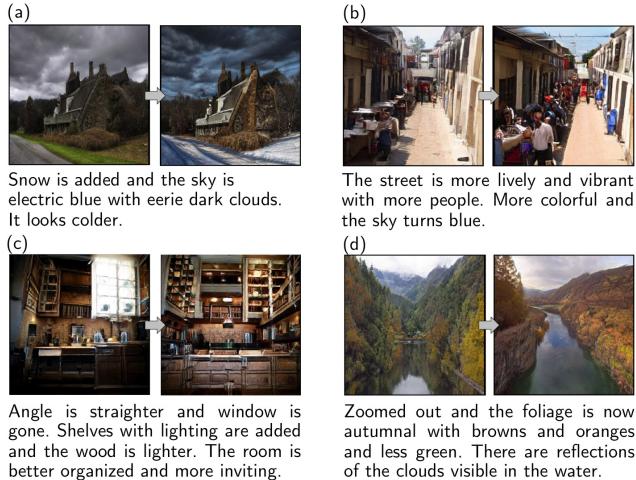


Figure 3: Sample transformations and AMT annotations from all four image classes: (a) cottage, (b) medina, (c) kitchen, (d) lake.

**Data normalization and post-processing.** To clean and normalize the direction annotations produced in Section 3.2, we first preprocess and lemmatize the labels using methods described in the supplement. Next, we post-

Image class	Distinct concepts	Repeated n>1 times	Unique to one class
<b>Cottage</b>	1166	508	147
<b>Kitchen</b>	1045	445	167
<b>Lake</b>	1167	479	153
<b>Medina</b>	1087	460	142
<b>All four</b>	2800	1372	609

Table 1: Distinct terms for concepts used in cleaned annotations, by class. We focus on those repeated in multiple labels, of which just under half (44%) appear in only one class.

process the labels by detecting phrases capturing decrease in a concept (e.g. *less green*, or *window is removed*), and assign them to individual negative directions. The result is a compact set of terms for human visual concepts describing each direction, which we refer to as *cleaned annotations*. For example, the cleaned annotation for the direction shown in Figure 3a would read “*snow, sky, electric, blue, eerie, dark, cloud, cold*.” Across all classes, 2800 unique concepts appeared, 1372 repeated at least once. 122 appeared in all four classes. The number of distinct concepts used in each class independently is shown in Table 1. Of concepts that appeared more than 20 times in the entire dataset (across all 4 classes), 32% are objects (e.g. *cabinet, tree*), 48% are attributes (e.g. *warmer, brighter*), and 20% describe scene- and object-level geometry (e.g. *background, angle*). We provide a more detailed description of these categories and a breakdown of concepts by image class in Section S.2.

The cleaned annotations indicate visual concepts that describe each of the LSDs. However, they do not isolate dimensions of variation that correspond to individual concepts; one direction may be described by multiple terms. To understand which visually salient terms can be mapped onto individual dimensions in the GAN’s representation, in Section 3.3 we disentangle the annotated directions into a set of principal perceptual components in the  $\mathcal{Z}$  latent space.

**Evaluating direction quality.** While our main contribution is a procedure (Section 3.3) for extracting a set of disentangled, human-recognizable concepts from *any* corpus of direction annotations, the method we describe for obtaining an initial set of directions also has advantages over related methods. To validate our decision to use the LSDs for annotation, we directly compared the annotations of LSDs in our dataset to two baselines: directions generated using the GANSpace method [10], and randomly generated directions. For a subset of 600 LSDs (150 in each of four image classes), we collected 10 annotations per direction using the AMT protocol described in Section 3.2. Additionally, we followed [10] and identified the same number of latent directions corresponding to principal components of feature

tensors on the first three layers of  $G$ . Finally, we sampled 600 random directions of fixed magnitude. All directions were normalized and added to the same set of  $\mathbf{z}$  with  $\alpha = 6$ .

Table 2 shows the results of our comparison. We find that LSDs elicit a more *diverse* vocabulary of both single-word concepts and their compositions. Additionally we measure inter-annotator BLEU [13] and inter-annotator BERTScore, where the latter leverages a large pretrained language model to measure semantic similarity between annotations [25]. While our LSDs obtain lower inter-annotator BLEU scores than the baselines, they obtain a larger BERTScore, suggesting there is less lexical overlap but greater semantic overlap in how annotators describe LSDs compared to the baselines. This hypothesis is further substantiated by the greater diversity of  $n$ -grams in LSD annotations.

Directions	1-grams	2-grams	3-grams	BLEU	BERTScore-R
Random	2,316	14,913	22,938	8.86	0.375
GANSpace	2,975	18,622	26,466	8.24	0.343
LSD (Ours)	3,156	20,986	31,307	7.17	0.393

Table 2: Comparison of diversity and reliability measures for 6000 annotations of directions added to the same set of  $\mathbf{z}$ . For the LSDs, observers recognize the most semantically similar changes per direction, and overall produce a larger number of both single-word concepts and their compositions.

### 3.3. Distilling directions for visual concepts

Our goal is to identify a vocabulary of *primitive* visual concepts, but as shown in Figure 3, the LSD annotations describe complex, compositional image changes, even after restricting annotation to layer-selective directions. We hypothesize that each annotated direction can be reconstructed from a set of *distilled directions* associated with individual concepts in the annotation. In other words,

$$d(\text{tall red building}) \approx d(\text{tall}) + d(\text{red}) + d(\text{building}) \quad (5)$$

This is a simplifying assumption (*red* suggests a different color in *red hair* vs *red brick*) [9]. However, it provides a convenient (and empirically effective) mathematical framework for distilling directions for primitive concepts from compositional annotations. In particular, we can formulate learning of the visual concept vocabulary as a regularized linear regression of the form:

$$\arg \min_{\mathbf{E}} \|\mathbf{W}\mathbf{E} - \mathbf{D}\|_F^2 + \lambda \|\mathbf{E}\|_F^2 \quad (6)$$

where rows  $i$  of word matrix  $\mathbf{W}$  correspond to annotations, and columns  $j$  of  $\mathbf{W}$  to individual words.  $\mathbf{W}_{i,j} = 1$  if word  $i$  appears in cleaned annotation  $j$ .  $\mathbf{W}\mathbf{E}$  is thus a matrix of annotation embeddings that we can compare to  $\mathbf{D}$ , where rows  $\mathbf{d}_i$  are the annotated directions in  $\mathcal{Z}$  latent space.

We may then solve analytically for  $\mathbf{E}$ :

$$\mathbf{E} = (\mathbf{W}^\top \mathbf{W} + \lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{D} \quad (7)$$

where  $\mathbf{I}$  is the identity matrix with the same size as  $\mathbf{W}^\top \mathbf{W}$ . The hyper-parameter  $\lambda$  determines the balance between the L2 loss and the regularization of  $\mathbf{E}$ . We set  $\lambda$  to 100 in our experiments.

The individual word embeddings  $\mathbf{e}_j$  in the latent space of  $G$  lie along the rows of  $\mathbf{E}$ . As in Section 3.1, transforming an image  $G(\mathbf{z})$  along the distilled direction corresponding to concept  $j$  is equivalent to moving in the direction  $\mathbf{e}_j$  in  $\mathcal{Z}$  latent space and passing the transformed  $\mathbf{z}$  vector through the generator:  $G(\mathbf{z} + \alpha \mathbf{e}_j)$ . The scaling parameter  $\alpha$  determines the degree and type of transformation: a larger  $\alpha$  introduces more of concept  $j$  to  $G(\mathbf{z})$ , and in many cases,  $-\alpha$  removes the visual concept from the scene. We note that the latent space is not perceptually uniform: steps of the same magnitude along different directions do not necessarily reflect the same amount of perceptual change. Continued work might map how this perceptual sensitivity to movement in each direction varies across the latent space.

Figures 1 and 4 illustrate the efficacy of applying our method to BigGAN-Places to disentangle directions corresponding to individual concepts, where each concept is associated with multiple annotated directions. We also tested the generalization of this approach to BigGAN-Imagenet, and show results in the supplement. Interestingly, *lake* is the only image class shared by both ImageNet and Places. For the same number of annotated directions (1280), the number of distinct concepts in the *lake* class for BigGAN-ImageNet is < 75% of the number of distinct concepts in the BigGAN-Places *lake* class. This could reflect less scene diversity in comparable ImageNet classes due to less training data. Given that our method is generalizable and fairly model-agnostic, we suggest that it could be used in such a manner to characterize a given generator by the projection of concepts salient to humans into the set of concepts the model has learned.

### 4. Evaluating distilled visual concepts

We have now distilled our LSDs into a vocabulary of primitive visual concepts, each consisting of a short language description, *e.g.* *snow* or *festive*, and an associated latent direction. Our next step is to evaluate how well the directions produce transformations in generated images that are faithful to their description. In other words, how often does adding the *trees* direction to a starting representation clearly add trees to the image?

We study this empirically by conducting a series of human experiments in which crowdworkers are asked to discriminate which among several image transformations corresponds to a specific visual concept. One of the transformed images is constructed by adding the corresponding

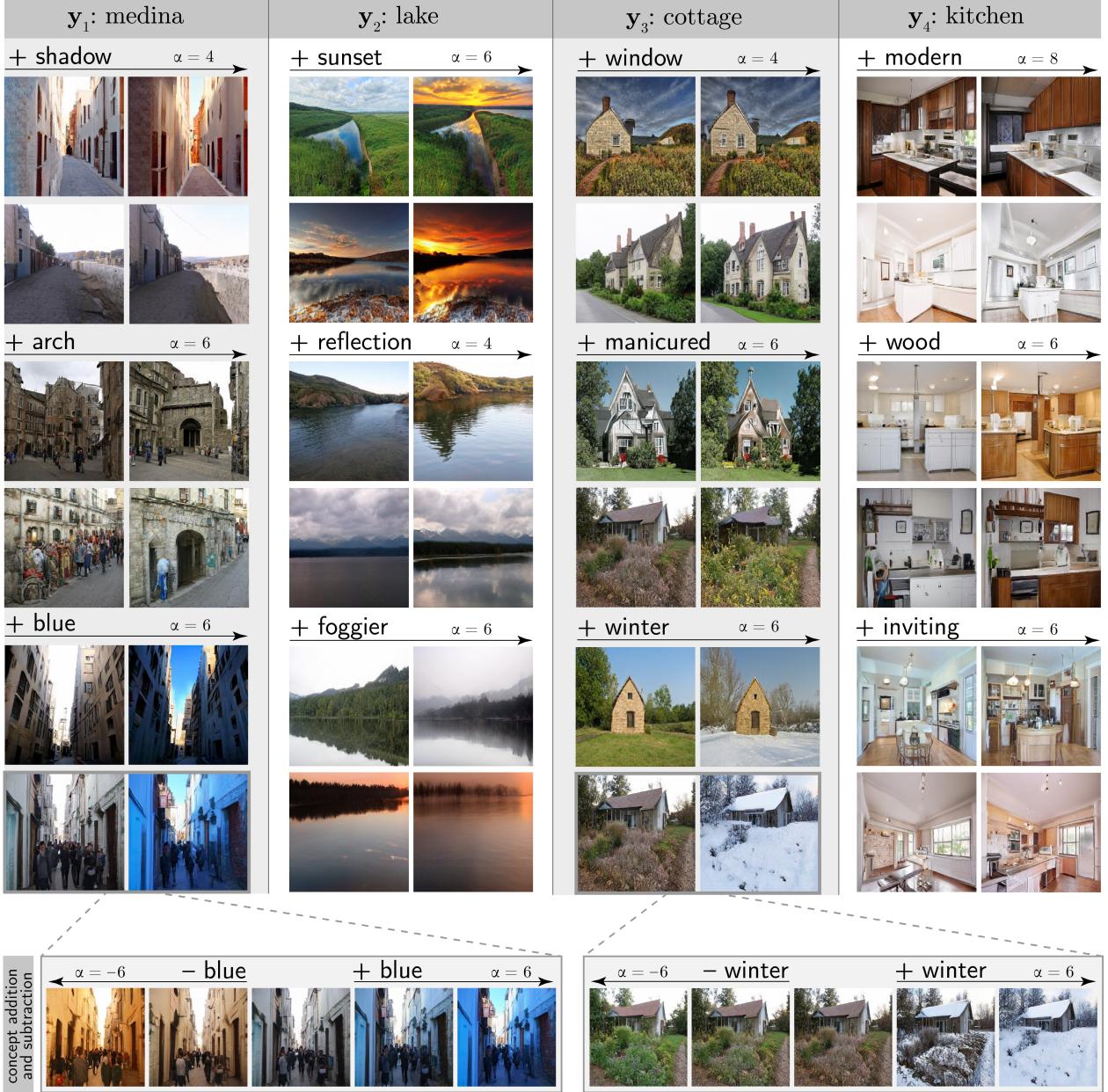


Figure 4: Example visual concepts across four classes of visual scenes, each applied to two  $\mathbf{z}$ . This sample represents only 13 of 1372 unique concepts discovered in BigGAN-Places. Some concepts (such as *blue*) occur in all scene classes. Others are characteristic of one or two (such as making a lake *foggier* or a kitchen *modern*). Bottom: in some cases, subtracting concepts can produce opposite transformations. For example, the subtraction of blue is the complementary orange, and the subtraction of winter is a spring scene. Additional examples with varying  $\alpha$  are shown in the supplement.

direction  $\mathbf{d}$  to the starting  $\mathbf{z}$ , while the others are constructed by adding *different directions from the vocabulary*. If humans reliably can discriminate which transformed image corresponds to the visual concept, that would suggest that the direction is faithful. The following three experiments adopt this structure and vary how the vocabulary is constructed in order to study different properties of the distilled

directions. The first two experiments focus on whether the directions **generalize** across starting representations (Section 4.1) and image classes (Section 4.2). The final experiment explores whether they reliably **compose** with one another, supporting combinatorial extensions to the vocabulary (Section 4.3).

## 4.1. Do concepts generalize across $\mathcal{Z}$ ?

We begin by asking whether distilled directions generalize to produce faithful transformations when added to unseen  $\mathbf{z} \in \mathcal{Z}$ , keeping all other inputs the same. Here, we fix a class  $\mathbf{y}$  and *only* vary the initial representation  $\mathbf{z}$ . This means that when we distil the vocabulary using Equation 7, we construct  $\mathbf{W}$  using only annotations for which the human annotator saw images generated with the class  $\mathbf{y}$ .

For each visual concept  $c_*$  and its distilled direction  $\mathbf{d}_*$ , we sample a  $\mathbf{z} \in \mathcal{Z}$  and three distractor directions  $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\}$  from the remaining directions in the vocabulary. Human participants are shown an initial image  $G(\mathbf{z}; \mathbf{y})$  and four transformed images  $G(\mathbf{z} + \alpha \mathbf{d}_i; \mathbf{y})$  for  $i = 1, 2, 3, *$  and are asked to discriminate which transformed image corresponds to  $c_*$ . If the direction  $\mathbf{d}_*$  successfully generalizes to the new  $\mathbf{z}$ , then participants should reliably choose the image change generated by that direction.

We recruit crowdworkers from Amazon Mechanical Turk; full details about the AMT setup and other hyperparameters can be found in the supplement. To denoise, we generate three sets of  $\mathbf{z}$ s and distractors per concept in the vocabulary, and additionally show each  $(\mathbf{z}, \mathbf{d})$  pair to five distinct participants, totaling 15 AMT HITs per concept.

**Distilled directions generalize to novel inputs.** Table 3 shows human accuracies by image class. Participants identify the correct image transformation more than 60% of the time, providing strong evidence that the distilled directions generalize across the representation space. Figure 5a shows that many concepts are recognized with higher accuracy than reported in Table 3, and only about 6% of concepts are recognized at the level of chance. *Attributes* are the most likely category of concepts to be accurately detected (75%). We include a further breakdown by concept in Section S.3.

**Detecting concepts with an SVM.** We replicated Experiment 1 using a linear classifier to detect concepts added to generated images, providing additional evidence that our vocabulary generalizes across  $\mathcal{Z}$ . For each of the 20 most frequent concepts in all four classes, we trained a linear SVM to distinguish the addition of that concept to an image from the addition of a randomly sampled distractor, and tested on held out images. Mean classification accuracy was significantly above chance in all classes (*cottage*: 80.2%, *kitchen*: 73.4%, *lake*: 79%, *medina*: 77.3%), and like humans, accuracy was highest overall for *attributes* (82.8%). We provide a per concept breakdown in the supplement.

## 4.2. Do concepts generalize across classes?

Visual concepts are context-sensitive. For example, making a kitchen scene *brighter* might involve adding additional light fixtures, while making a cottage scene brighter

Experiment	Kitchen	Lake	Medina	Cottage	Avg.
Generalize $\mathbf{z}$	.60	.76	.62	.64	.66
Generalize $\mathbf{y}$	.37	.39	.43	.37	.39
Composition	.40	.44	.51	.41	.44

Table 3: Human accuracy discriminating a target concept from three distractors, where the concept is visualized by applying its associated direction to new  $\mathbf{z}$  (Generalize  $\mathbf{z}$ ), new classes (Generalize  $\mathbf{y}$ ), and compositions of directions (Composition).

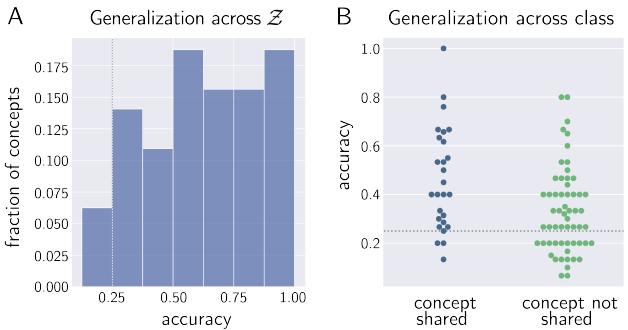


Figure 5: (a) Histogram from Section 4.1, where *accuracy* refers to the fraction of times that humans correctly recognized a specific concept. Dotted vertical line demarcates accuracy of random guessing. For 94% of concepts, participants recognize the correct change more often than if they guessed randomly, suggesting that the directions generalize across  $\mathbf{z}$ . (b) Concept accuracies from the cross-class evaluation of Section 4.2, bucketed by whether the concept appeared in both annotations for the training class and the test class. Some concepts (typically objects and attributes) exhibit strong cross-class generalization, with one being correctly recognized by every observer. Other concepts fail to generalize even when they appear in annotations for both classes, suggesting BigGAN has not entirely disentangled concept from class.

will likely involve intensifying the sun. Despite the differences between these image transformations, both are instantiations of the visual concept *brighter*. At the same time, some visual concepts might be unique to a context. The kitchen class exclusively features concepts like *cabinets* and *appliances*, while the lake class features *snow* and *mountains*. This raises the question: if we construct a vocabulary using annotations from one image class, do the resulting directions produce faithful transformations on other classes?

We now repeat our evaluation from Section 4.1, but instead of fixing  $\mathbf{y}$  in the evaluation, we choose it at random from the set of classes not used to construct  $\mathbf{E}$  in Equation 7. Hence, when evaluating the *kitchen* vocabulary, we generate images and transformations with the *lake*, *cottage*, or *medina* class. We draw several conclusions.

**Generalization across class is most robust when concepts are shared between classes.** Figure 5b shows that participants recognize concepts most often when the concept appears in the vocabulary for both classes. This agrees with the intuition that it should be difficult to add a visual

concept to image when that concept is foreign to the context, *e.g.* adding *appliances* to a lake scene. For these transformations to succeed, BigGAN would have to generate out-of-distribution images.

**However, distilled directions still generalize across classes.** Even though cross-class generalization is harder than within-class generalization, humans still recognize the target visual concept a majority of the time. This even includes some out-of-distribution generalizations like the one shown in Figure 6, which inserts snow into a medina, despite snow being unseen in medina training images.

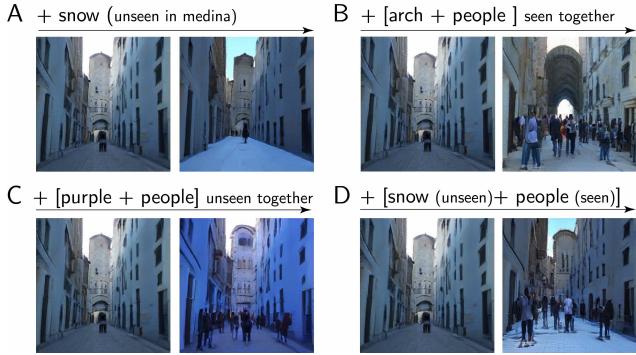


Figure 6: Several image changes produced by decomposed directions applied to the same starting image of a medina. Directions generalize (a) across and (b) within class. Two directions can be composed regardless of whether the corresponding concepts (c) did not co-occur in the original corpus or (d) did co-occur.

### 4.3. Do concepts compose?

In the previous experiments, our vocabulary consisted of primitive visual concepts such as *mountain* and *dark*. Can we construct more complex visual concepts from these primitive ones? One way to do this would be to compose the primitive concepts conjunctively: given a *mountain* direction and a *dark* direction, construct a *mountain*  $\wedge$  *dark* by simply averaging the two directions.

Our goal in this section evaluate how often composition of this kind succeeds. We repeat the evaluation from Section 4.1, now constructing the vocabulary by conjunctively composing every pair of primitive concepts from the original vocabulary. Formally, given a primitive vocabulary  $V$  for a fixed concept  $y$  and two directions  $a, b \in V$ , we define their composition  $a \circ b$  to be  $(a + b)/2$  and define our new vocabulary to be  $V' = \{a \circ b : (a, b) \in V^2\}$ . In practice,  $V'$  is quite large because it has quadratically many concepts, so we select a random subset of 50 compositions.

As before, for each direction  $a \circ b \in V'$ , we sample a representation  $z \in \mathcal{Z}$  and three distractor directions. However, now we choose two of the distractors to be compositions of  $a$  and  $b$  with other primitives. Specifically, we sample

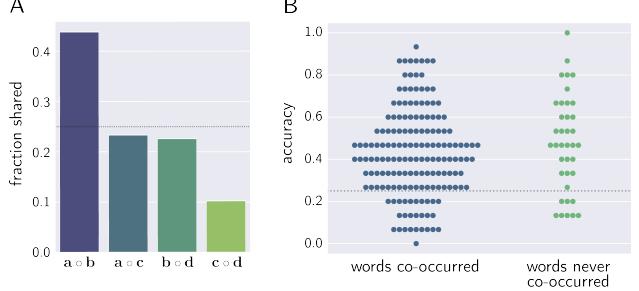


Figure 7: (a) Fraction of times humans chose each composition in Section 4.3.  $a$  and  $b$  are target directions, while  $c$  and  $d$  are randomly chosen distractors. Observers frequently recognize the correct composition, but even when not, they prefer partially correct compositions, suggesting the decomposed directions compose faithfully. (b) Fraction of times humans recognized each concept composition, bucketed by whether composed concepts co-occurred in the original corpus. Both classes of composition perform have comparable mean accuracies, suggesting many of the directions in the vocabulary can be faithfully composed.

two additional directions  $c, d \in V - \{a, b\}$  and use  $a \circ c$ ,  $b \circ d$ , and  $c \circ d$  as distractors. Participants then discriminate which transformed image contains both  $a$  and  $b$ .

**Distilled directions compose to produce new and recognizable concepts.** Even though compositional changes are harder to discriminate, participants still predict the correct change reliably above chance. Furthermore, Figure 7a shows that when participants choose a distractor, they tend to pick distractors closest to the target, i.e.  $a \circ c$  or  $b \circ d$ .

**Composition produces faithful transformations even when concepts did not co-occur in the training data.** Figure 7b shows that participants recognize composed concepts regardless of whether the constituent concepts ever appeared together in a single LSD description. Figure 6 shows an example, in which the *purple* and *people* concepts (unseen together during training) can be composed to produce an image of a purple medina filled with people.

## 5. Conclusion

We introduce a new procedure for building open-ended vocabularies of primitive visual concepts represented in GANs’ latent spaces, and show that these concepts are reliably recognizable and freely composable. This work represents an important step toward bridging the representational gap between human perception and artificial generators. Future work could explore the use of our approach with generators other than BigGAN, such as StyleGAN.

**Acknowledgements.** We thank the MIT-IBM Watson AI Lab for support, and IBM for the donation of the Satori supercomputer that enabled training BigGAN on MIT Places. We also thank Luke Hewitt for valuable discussion and insight.

## References

- [1] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics*, 38(4):1–11, Jul 2019. [2](#)
- [2] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks, 2018. [2, 3](#)
- [3] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate, 2019. [2](#)
- [4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. [3](#)
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. [3](#)
- [6] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. Image counterfactual sensitivity analysis for detecting unintended bias, 2020. [2](#)
- [7] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties, 2019. [1, 2, 3](#)
- [8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. [1](#)
- [9] Irene Heim and Angelika Kratzer. *Semantics in Generative Grammar*. Blackwell, 1998. [5](#)
- [10] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls, 2020. [1, 2, 4](#)
- [11] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks, 2020. [1, 2, 3](#)
- [12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. [2](#)
- [13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. [5](#)
- [14] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019. [2](#)
- [15] William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement, 2020. [1, 2](#)
- [16] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations, 2020. [2](#)
- [17] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. [2](#)
- [18] Sarah Schwettmann, Hendrik Strobelt, and Mauro Martino. Latent compass: Creation by navigation, 2020. [3](#)
- [19] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing, 2020. [1, 2](#)
- [20] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans, 2020. [1](#)
- [21] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space, 2020. [1, 2](#)
- [22] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation, 2020. [2](#)
- [23] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding, 2018. [2](#)
- [24] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis, 2020. [1, 2, 3](#)
- [25] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. [5](#)
- [26] Bolei Zhou, Ágata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 07 2017. [3](#)
- [27] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold, 2018. [2](#)