

Interpretable Transformations with Encoder-Decoder Networks

Daniel E. Worrall

Stephan J. Garbin

Daniyar Turmukhambetov

Gabriel J. Brostow

University College London *

Abstract

Deep feature spaces have the capacity to encode complex transformations of their input data. However, understanding the relative feature-space relationship between two transformed encoded images is difficult. For instance, what is the relative feature space relationship between two rotated images? What is decoded when we interpolate in feature space? Ideally, we want to disentangle confounding factors, such as pose, appearance, and illumination, from object identity. Disentangling these is difficult because they interact in very nonlinear ways. We propose a simple method to construct a deep feature space, with explicitly disentangled representations of several known transformations. A person or algorithm can then manipulate the disentangled representation, for example, to re-render an image with explicit control over parameterized degrees of freedom. The feature space is constructed using a transforming encoder-decoder network with a custom feature transform layer, acting on the hidden representations. We demonstrate the advantages of explicit disentangling on a variety of datasets and transformations, and as an aid for traditional tasks, such as classification.

1. Introduction

We seek to understand and exploit the deep feature-space relationship between images and their transformed versions. Different feature spaces are illustrated in Figure 1, and support different use-cases: separability helps discriminate between categories such as identity, while invariance improves robustness to nuisance variables during data capture. Taking head pose as an example, what is a nuisance for one task could be the focus of another. Therefore, we propose deep features with transformation-specific *interpretability*, which combine both (1) discriminative and (2) robustness properties, with the further benefits of (3) a user-guided parameterized space for controlling image synthesis through interpolation.

Learning such a feature space is difficult. In image data, transformations of objects usually couple in complex nonlinear ways, leading to an *entangling* of transformations. The reverse process of *disentangling* is then especially hard. An obvious post hoc

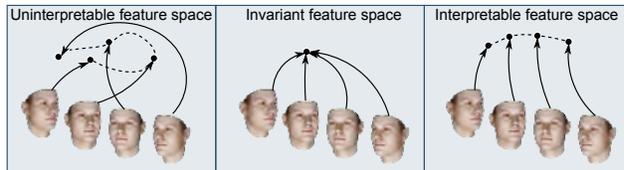


Figure 1. Three alternative feature spaces and how each encodes images of the same person. (Left) A feature space that is hard to interpret, similar to one learned by a typical CNN. While transformation information is present, it is not obvious how to extract that directly from the feature space. (Middle) A transformation-*invariant* feature space. (Right) An interpretable feature-space, where ordered transformations of the input subject relate to ordered, structured features. This is like a learned metric space, but also allows for image synthesis. Images of another person are not shown, but would ideally project similarly, albeit elsewhere in each feature space.

solution is to learn disentangling transformations using a regressor [31], but this is a time-consuming and inexact process. We cannot assume that the change in representation of a chair and its rotated twin is necessarily the same as the change in representation between a banana and its equally rotated twin. We propose disentangling as an end-to-end supervised learning problem. Some image variations are hard to quantify or explain. But others, for instance 2D and 3D warps or color appearance changes, allow ready access to pre- and post-warp image pairs, along with their ground-truth transformation parameters. These easier transformations, we find, lend themselves to smooth parameterization in feature space, and therefore interpretability. One could argue that it is nicer to learn everything only from raw data, but the transformation parameter labels considered here are obtained with little or no human effort. We therefore pre-define the feature-space structures that encode basic transformations, and train neural networks that map into and out of this feature-space.

We take our motivation from considering the feature space structure, introduced by convolutional neural networks [30] (CNNs). CNNs owe their success to two differences from the older and more general multilayer perceptrons [36]: 1) the receptive field of deep neurons is localized to a small neighborhood, typically not greater than 7×7 pixels from the layer below, and 2) incoming weights are tied between all translated neurons. The motivation behind translational weight-tying is that correlations in the activations are invariant under translation. The side-effect of enforcing such a structure on the weights of a neural network

*<http://visual.cs.ucl.ac.uk/pubs/interpTransform/>

is that integer pixel translations of the image input induce proportional integer pixel translations of the deep feature maps. This phenomenon is called *equivariance*, meaning the feature-representation of a shifted input is the same, save for its location. We explore continuous transformation equivariance for CNNs, and for the first time, for fully connected models.

In this paper, we consider rotations in 2D and 3D, out-of-plane rotations, small translations, stretchings, uniform scalings and changes in lighting direction. For these transformations CNNs do not generally display the equivariance property; although, there are a number of works, which do tackle the problem of rotation [6, 10, 41, 12, 16, 28, 50, 15, 56]. The main problem with all these approaches (which we detail in the next section) is that the equivariance properties are handcrafted, and suffer from unmodeled oversights in the design process. For instance, all but [50] consider equivariance to discretely sampled rotations, when real world rotations are in fact continuous. Given that we can simulate many image-space transformations, it seems only natural to simply acquire equivariance through learning.

We now cover related work and theory, followed by Section 3 where we introduce our method and the new *feature transform layer*, and Section 4 where we test our framework on de-render-re-render problems and for view independent features.

2. Related Work and Theory

Here we outline basic concepts for us to formalize the task of encoding interpretable transformations, and break down a list of related works into categories of handcrafted or learned equivariance in traditional vision and deep learning.

Definition 1 A function $\mathbf{f} : \mathcal{X} \rightarrow \mathcal{Y}$ is equivariant [49] under a set of transformations Θ if for any transformation $\mathcal{T} : \Theta \times \mathcal{X} \rightarrow \mathcal{X}$ of the input, we can associate a transformation $\mathcal{F} : \Theta \times \mathcal{Y} \rightarrow \mathcal{Y}$ of the output such that

$$\mathcal{F}_\theta[\mathbf{f}(\mathbf{x})] = \mathbf{f}(\mathcal{T}_\theta[\mathbf{x}]), \quad (1)$$

for all $\theta \in \Theta$. Transformations \mathcal{T}_θ and \mathcal{F}_θ represent the same underlying transformation but in different spaces, denoted θ .

Equivariance is desirable, because it reveals to us a direct relationship between image-space and feature-space transformations, which for deep neural networks are usually elusive [31]. Note that *invariance* is a special case of equivariance, where $\mathcal{F}_\theta = \mathbb{I}$ is the identity for all input transformations.

Definition 2 We define an interpretably equivariant feature-space to be an equivariant feature-space as in Equation 1, where the transformation functions \mathcal{F}_θ and \mathcal{T}_θ are quantitatively known and can be implemented for all θ , \mathbf{x} and \mathbf{f} .

At an abstract level, an equivariant function is one where some level of structure is preserved between the input and output. Interpretability is the added requirement that for a given θ we know how to apply \mathcal{F}_θ and \mathcal{T}_θ . It may be the case that one of

these transformations is complicated and cannot be written down as a mathematical expression in closed form (e.g., the rendering equation), but as long we are able to simulate it that is enough. As we show in Section 3.2, one way of preserving the structure of transformations across a feature mapping is via a condition called the *homomorphism property*. In all of the subsequent related works, equivariance to transformations is the central theme.

Handcrafted methods In the 1980s, Crowley and Parker [9] studied scale-space representations. These are formed by convolving images with scaled versions of a filter. Scale-space methods exhibit interpretable equivariance. They can be extended to invertible transformations by transforming the filters [35, 1] but has computational complexity exponential in the number of degrees of freedom (DOF) of the transformation. Furthermore, we can only convolve with a finite number of filters, when in reality many transformations are continuous. Freeman and Adelson [13] and Lenz [33] simultaneously solved the continuity problem, through orientation steerable filters w_θ . These can be synthesized at any continuous orientation θ . These are formed as a linear combination of fixed basis filters ϕ_n :

$$w_\theta(\mathbf{x}) = \sum_{n=1}^N \alpha_n(\theta) \phi(\mathbf{x}). \quad (2)$$

$\alpha_n(\theta)$ are known as the *interpolation functions*. These are still band-limited but unlike scale-space the frequency characteristics are easier to design. Steerable filters were extended to most transformations with one DOF (one-parameter subgroups) [47, 45], for instance, 1D translations, 2D rotations, scalings, shears, and stretches. For these transformations, there is a function ρ , under which transformation θ becomes a shift, so $I(x) \xrightarrow{\mathcal{T}_\theta} I(\rho^{-1}(\rho(x) - t_\theta))$, where t_θ is the shift. Meanwhile, Perona [42] showed that in practical situations some transformations cannot be enacted exactly using steerable functions, for instance scale and affine transformations (specifically those which do not have compact group structure). He showed these can be approximated well with very few basis functions, computed from the singular value decomposition of a matrix of transformed versions of a template patch. This is limited by template choice, SVD efficiency, and figuring out the interpolation functions for steering. More recently Hasegawa [17] and Koutaki [26] used a variant of this method to learn an affine-equivariant feature detector.

Invariance to 1 DOF transformations can be gained via the Fourier Transform (FT) Modulus method [25]. This uses the time-shifting property of the FT $w(x-t) \xleftrightarrow{FT} e^{i\omega t} \mathcal{W}(\omega)$, where $\mathcal{W}(\omega)$ is the FT of $w(x)$. The FT modulus $|e^{i\omega t} \mathcal{W}(\omega)| = |\mathcal{W}(\omega)|$ is independent of the shift t . As noted in Scattering Networks [4], this operation removes excessive localization information and is unstable to high-frequency deformations noise. They instead take the modulus of the response to a bank of discretely rotated and scaled wavelets, repeatedly in a deep fashion. This is perhaps the most successful version of a handcrafted deep equivariant feature map.

Neural Networks Equivariance in deep learning has very deep roots as far back as the early 1990s. Barnard and Casasent [2] split the main approaches to transformation invariance into three categories: 1) *Data augmentation*: This is effective and simple to implement, but lacks interpretability. 2) *Preprocessing*: This is effective, but cannot be applied to geometric transformations. 3) *Structured weight networks*: These are numerous in the literature. CNNs [30] are the most famous example. Pixel-wise integer shifts of an input image will induce proportional pixel-wise shifts in the deep feature space. For partial translation invariance, there is the Global Average Pooling layer [34]. For rotations there are two major approaches for discrete rotations: rotate the filters [6, 8, 16, 41, 15, 56] and rotate the input/feature maps [10, 12, 28]. Continuous rotations were recently proposed by [50]. They restrict their filters and architectures so that the convolutional response is equivariant to continuously rotated inputs. Beyond rotation, [18] warp the input, so that general transformations are globally linearized, facilitating the application of CNNs. This requires prior knowledge of the type of transformation and where it is applied in the image. [8] can deal with multiple transformations, but these are restricted to group-theoretic structures. [22] are able to explicitly transform feature maps with the spatial transformer layer, but do not transform features in the channel dimension. In contrast to the above methods, our method is general and does not require extensive architectural engineering. We can also disentangle confounding factors such as out-of-plane rotation and lighting direction.

Deeply Learned Equivariance Some have sought to learn equivariance directly from data. These broadly split into purely generative, purely discriminative and auto-encoded methods. **Discriminative**: [32] regress affine equivariant feature-descriptors directly using supervised data. Their framework is easy to implement, but restricted to group-theoretic transformations. **Generative**: [11] generate views of 3D chairs by regressing appearance with a CNN from an embedding space. In InfoGAN, [5] instead used a mutual information maximizing criterion for unsupervised learning of the ‘natural’ transformations in a training set. This mostly manages to disentangle transformation, but unlike [11] is non-interpretable. **Auto-encoded**: [27] presented the deep convolutional inverse graphics network (DC-IGN), a partially supervised variational auto-encoder [24], equivariant to out-of-plane rotation and relighting. Their model is impressive but requires a complicated training procedure, is partially interpretable, and unlike us does not fully exploit known supervised information about transformations. [39, 19, 55] instead reconstruct transformed versions of an image, given the image and transformation parameters as input. These are similar to our method, but cannot be used to extract interpretable transformation equivariants, which we can do. [7] does learn interpretable equivariance to manipulate images of 3D objects from 2D images, but this is only demonstrated on 3D rotations. [43] also does learn interpretable equivariance for 3D volumes from 2D images, but their

representation space is entire 3D volumes. This is impressive, but it is computationally expensive to represent entire volumes in memory, when sometimes it may not be necessary.

3. Method

CNNs are interpretably equivariant to pixel-wise translations of their input up to boundary effects, but not to transformations such as 2D and out-of-plane rotations, uniform scalings, stretches, relighting, flips, etc. In this section we design a neural network to learn an interpretable transformation equivariant feature-space. Our method can cope with continuous transformations on intervals, for example, uniform scalings and stretches, and continuous transformations on circles, such as, geometric rotation and relighting, but not discrete transformations, like vertical flips. In Section 3.1 we outline our general framework and in Section 3.2 we introduce the *feature transform layer*, a channel-wise analogue of the spatial transformer, which can also be applied to fully-connected layers.

3.1. Problem Setup

We assume that we are given a training set $\mathcal{D} = \{(\mathbf{x}^1, \tilde{\mathbf{x}}_{\theta^1}^1, \theta^1), \dots, (\mathbf{x}^N, \tilde{\mathbf{x}}_{\theta^N}^N, \theta^N)\}$ containing pairs of views of transformed examples $(\mathbf{x}^i, \tilde{\mathbf{x}}_{\theta^i}^i)$ and relative transformation vectors θ^i . The relative transformations may be the result of a sensor measurement, or they may be the result of artificial data augmentation, in which case the training set is potentially infinite. The task is to predict $\tilde{\mathbf{x}}_{\theta^i}^i$ given \mathbf{x}^i and θ^i (from now on we just write θ for short). We use relative transformation information instead of absolute transformations, because there is no canonical pose, which generalizes across object classes, where alignment between, say, a banana and an airplane does not make sense.

Many images $\mathbf{x} \in \mathcal{X}$ are formed from capturing an object $\mathbf{o} \in \mathcal{O}$ in the 3D world projected via a function $\Pi: \mathcal{O} \rightarrow \mathcal{X}$ onto a 2D canvas. To transform image \mathbf{x} into $\tilde{\mathbf{x}}_{\theta}$ we have to invert Π to find \mathbf{o} , perform the world-space transformation and re-project back into image space, so

$$\tilde{\mathbf{x}}_{\theta} = \Pi[\mathcal{T}_{\theta}[\mathbf{o}]] = \Pi[\mathcal{T}_{\theta}[\Pi^{-1}[\mathbf{x}]]]. \quad (3)$$

The problem with this approach is that Π is in usually non-invertible. Our solution is to infer the 3D object \mathbf{o} given \mathbf{x} via statistical methods. CNNs are good at this kind of task (e.g., [27]), so we opt to use a CNN. Now storing a full volumetric representation like in [43] is costly, so we instead opt to use a compressed feature encoding $\mathbf{e}(\mathbf{x})$ to approximately represent \mathbf{o} , this requires we also have a feature-space representation of the transformation, \mathcal{F}_{θ} —see Section 3.2 for details. In our case the feature space is partially learnable, with pre-defined structure imposed by \mathcal{F}_{θ} . Our basic model is shown in Figure 2, it is an encoder-decoder network. Loosely speaking

- $\mathbf{e}(\bullet)$ approximates $\Pi^{-1}[\bullet]$,
- \mathcal{F}_{θ} is the feature space equivalent to \mathcal{T}_{θ} ,
- $\mathbf{d}(\bullet)$ approximates $\Pi[\bullet]$,

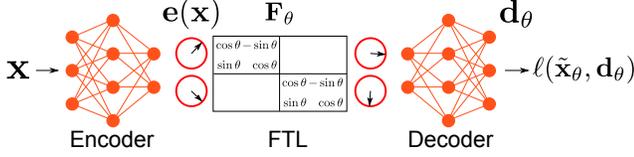


Figure 2. We enforce equivariance by minimizing the loss ℓ between reconstruction of transformed features \mathbf{d}_θ and a transformed target $\tilde{\mathbf{x}}_\theta$. Given just \mathbf{x} , the encoder-decoder network does not have enough information to produce a transformed output, thus supplying the missing information θ via the feature transform layer (FTL) forces the network to learn a mapping in and out of the FTL. Critically, whereas other approaches, such as transforming auto-encoders [19] and InfoGAN [5], learn the reconstruction to be sensitive to feature transformation information, we can simultaneously learn to map from images to transformation equivariant features.

where we have written $\Pi^{-1}[\bullet]$ to mean inversion of the projection if possible, or approximation of it. We train the weights of the encoder and decoder by minimizing a summed reconstruction loss ℓ , where

$$\mathcal{L}(\mathcal{D}) = \sum_i \ell(\mathbf{d}(\mathcal{F}_\theta^i[\mathbf{e}(\mathbf{x}^i)]), \tilde{\mathbf{x}}_\theta^i). \quad (4)$$

In our experiments we use a diverse set of losses, namely, L1 loss, SSIM, and balanced cross-entropy. Note that since we define \mathcal{F}_θ the feature space of encodings $\mathbf{e}(\mathbf{x})$ is interpretable by Definition 2. In Section 3.2, we demonstrate an encoding, which enforces explicit disentangling and from which we can gain approximate transformation invariance ‘for free’.

3.2. The Feature Transform Layer

The feature-space equivalent of the image-space transform \mathcal{T}_θ is the *feature transform layer* \mathcal{F}_θ . It is an analogue of the spatial transformer [22], but applied to general feature-spaces, not necessarily with spatial dimensions. This means that we can apply it to fully connected layers as well as convolutional layers. It is easiest to describe the feature transform layer via its implementation.

Consider a feature vector \mathbf{e} , which may be a column of CNN feature channels above a pixel location in an image, or the output of a fully-connected layer. The feature transform layer performs a linear transformation of \mathbf{e} via matrix \mathbf{F}_θ , such that the output \mathbf{y} of the layer is

$$\mathbf{y} = \mathcal{F}_\theta[\mathbf{e}] = \mathbf{F}_\theta \mathbf{e}. \quad (5)$$

We only consider linear transformations, where

$$\mathbf{F}_{\theta_2 \theta_1} = \mathbf{F}_{\theta_2} \mathbf{F}_{\theta_1}. \quad (6)$$

This condition says that if we apply transformation θ_1 to an image, followed by transformation θ_2 , which we have written as $\theta_2 \theta_1$, then in feature space this should be equivalent to applying \mathbf{F}_{θ_1} followed by \mathbf{F}_{θ_2} . We refer to Equation 6 as the

homomorphism property. Abstractly, we can think about it as forcing the neural network to learn a mapping from image-space to feature-space, which preserves the intrinsic structure of the transformations. The homomorphism property implies that (see Supplementary Material)

$$\mathbf{F}_{\theta_1^{-1}} = \mathbf{F}_{\theta_1}^{-1}. \quad (7)$$

This means that invertible transformations of the input are invertible in feature-space. The homomorphism property is key to ensuring that transformation information is not lost when mapping into feature-space. Examples of \mathbf{F}_θ are N -dimensional rotation matrices, also known as $SO(N)$, full-rank diagonal matrices, or most generally the group of invertible $N \times N$ matrices, known as $GL(N)$. We use rotation matrices, \mathbf{R}_θ , which have the additional property of being orthogonal or *norm-preserving*. This means that we can use the feature vector lengths as transformation invariants because

$$\|\mathbf{R}_\theta \mathbf{e}\|_2^2 = \mathbf{e}^\top \mathbf{R}_\theta^\top \mathbf{R}_\theta \mathbf{e} = \mathbf{e}^\top \mathbf{e} = \|\mathbf{e}\|_2^2, \quad (8)$$

which shows that $\|\mathbf{R}_\theta \mathbf{e}\|_2^2$ is in fact independent of θ . Feature vectors are usually high-dimensional consisting of many channels. We therefore implement the feature transform layer by applying the same rotation matrix on multiple groupings of channels, which we call *subvectors* of \mathbf{e} . We can then define a larger set of invariants, by measuring the relative phase between different subvectors. These are invariant to θ , because if \mathbf{e}_1 and \mathbf{e}_2 are two subvectors of \mathbf{e} , then

$$(\mathbf{R}_\theta \mathbf{e}_2)^\top \mathbf{R}_\theta \mathbf{e}_1 = \mathbf{e}_2^\top \mathbf{R}_\theta^\top \mathbf{R}_\theta \mathbf{e}_1 = \mathbf{e}_2^\top \mathbf{e}_1, \quad (9)$$

which is independent of θ . If $\mathbf{e}_1 = \mathbf{e}_2$, this reduces down to the feature vector length. We denote the concatenation of all subvector dot products as $\|\mathbf{e}\|_{\mathcal{F}}$. While at first not obvious, we can encode many transformations using rotation matrices, even ones which do not have periodic structure. The trick is to map the domain of the transformation onto the half-circle/sphere, see Figure 3. We prefer to do this rather than using another, perhaps more natural, representation because of the convenience of taking L2-norms and inner products to form invariants.

Disentangling We now consider how to disentangle transformations. Since we can model transformations, whose

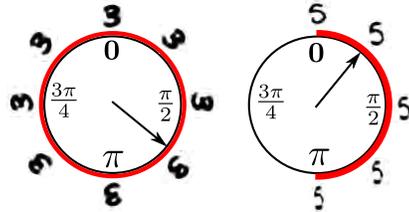


Figure 3. We encode transformations by mapping them on to circles and N -dimensional hyperspheres in feature space. This parameterization can deal with periodic and bounded transformations on an interval. The L2-norm of the result feature vectors are transformation invariant.

Method	$\tilde{\mathbf{x}}_\theta \theta$	$\tilde{\mathbf{x}}_\theta \theta,\mathbf{x}$	$\theta \mathbf{x}$	CNN	MLP	Interpretable	Supervised	Image size
DC-IGN [27]	✓	✗	*	✓	✓	†	‡	150x150
InfoGAN [5]	✓	✗	✗	✓	✓	✗	✗	64x64
Generating Chairs [11]	✓	✗	✗	✓	✓	✓	✓	128x128
Transforming AEs [19]	✗	✓	✗	✗	✓	✗	✓	96x96
Learned Visual Reps. [7]	✗	✓	✗	✗	✓	✓	✗	96x96
Unsup. 3D from images [43]	✗	✓	✗	✓	✓	✓	✗	30x30x30
Covariant features [32]	✗	✗	✓	✓	✓	✓	✓	57x57
Spatial Transformer [22]	✗	✓	✗	✓	✗	✓	-	Any
Ours	✗	✓	✓	✓	✓	✓	✓	150x150

Table 1. Comparison of method scopes. In the first 3 columns we display whether a method can generate an image $\tilde{\mathbf{x}}_\theta$ given just parameters θ , $\tilde{\mathbf{x}}_\theta|\theta$; conditioned on an original image $\tilde{\mathbf{x}}_\theta|\theta,\mathbf{x}$; or infers transformation parameters given an image $\theta|\mathbf{x}$. * Qualitative relationship only. † Correspondence between feature dimensions and transformations known, qualitative relationship only. ‡ Partial supervision: minibatches grouped into variation of single parameter, but values not given.

parameters exist on a circle or interval, we can model each independent transformation DOF by mapping it to a different circle or half-circle. Some transformations, like lighting direction, are more conveniently mapped to the surface of a 3D-sphere. Thus the feature transform layer is

$$\mathbf{F}_\theta \mathbf{e} = \begin{bmatrix} \mathbf{R}_{\theta_1} & & \\ & \ddots & \\ & & \mathbf{R}_{\theta_N} \end{bmatrix} \mathbf{e}, \quad (10)$$

with possible tied θ_i when we apply a transformation to multiple subvectors. The feature transform layer is simple to implement—it is just a matrix multiplication and the block diagonal structure allows efficiency saving via reshapes. In our experiments we found a slow down of just 2%. Furthermore, it can be applied to convolutional features in synchrony with a spatial transformer [22] for complete control of both spatial and feature properties.

4. Experiments, Results, and Discussion

Below we demonstrate the ability of our system to learn meaningful features on MNIST [54], MNIST-rot [29], the Basel Face Dataset [20], and ModelNet10 [51]. We choose these datasets because they demonstrate our system’s general-purpose usage and performance on 2D and 3D images, for transformations with complex entanglement, and with and without information loss. Our encoder-decoder structure is shown in Figure 5. They are all implemented in TensorFlow.

4.1. MNIST: 2D images—2D transformations

This experiment demonstrates our system’s ability to disentangle confounding transformations and how it reconstructs an input, after manipulation of the features. The MNIST dataset [54] contains 50k training, 10k validation, and 10k grayscale test images of handwritten digits, size 28×28 . The images are very simple, usually just a pen-stroke. We apply random scalings in the x- and y-directions followed by a random 2D rotation. Due to the simplicity of the images, we use an MLP

for both encoder and decoder. Both encoder and decoder have 3 layers, separated by batch normalization [21] and leaky ReLU nonlinearities [37] apart from the input and output of the feature transform layer, which are linear. All layers except the input and output are 510 neurons wide¹. The feature transform matrices are a block diagonal composition of three 2D rotation matrices repeated 85 times: rotation \mathbf{R}_{rot} , x-scaling $\mathbf{R}_{\text{scale-x}}$, and y-scaling $\mathbf{R}_{\text{scale-y}}$. We train with the Adam optimizer [23] for 200 epochs, with minibatch size 128 and initial learning rate 10^{-3} . After training we pass a random digit from the test set through the encoder and transform the code by multiplying by feature transform matrix \mathbf{F}_θ . In Figure 4 we show random digits from the test set, slowly varying the transformation vectors on an interval. Each row shows a random digit under a combination of rotation, x-, and y-scaling. Notice how the encoder-decoder successfully learns to rotate digits, solely from the feature transformation. Notice also that the scalings are applied in the x- and y-directions of a coordinate system aligned to the canonical pose

¹We use this non-standard width because we model three transformations, with each transformation modeled on a separate circle. So feature-space dimensionality must be a multiple of $3 \times 2 = 6$. Furthermore, the value of 510 is close to 512, a common feature-space dimensionality.



Figure 4. MNIST reconstructions: The left most column indicates transformation. The second to left column shows the input. Subsequent columns show the transformed images. The reconstruction struggles slightly with enlarged images, on the left, but on the whole clearly show that we have control over the disentangled representation. Notice that the x- and y-scalings are in the coordinate frame of the canonical pose of the digits. This demonstrates the ability to disentangle confounding transformations.



Figure 6. (Viewed best in color). Relit and re-rotated reconstructions from our Basel face encoder-decoder model. The input faces shown on the left (blue box) are not presented to the encoder-decoder at training time. From left to right we vary one degree of freedom only. Outside the large green box the encoder-decoder has never seen those transformation parameters. We note the impressive ability of the model to rotate out-of-plane and to relight a 3D surface, when only given a 2D input and a pair of 3D rotation matrices. For unseen transformation parameters, notice that the relighting is of perceptually decent quality, but that the geometric rotations degenerate in quality around the boundaries, such as the nose and chin.

We dub it the *facial transformer*. As basic design principles, we avoid max-pooling, favoring strides, and use batch normalization and leaky ReLUs after all layers, apart from before and after the feature transform layer. For deconvolution we upsample with nearest-neighbor interpolation followed by regular convolution [40, 27]. Inspired by [53, 14] our reconstruction loss is a convex combination of the structural similarity index (SSIM) [48] and L1 loss. The L1 loss encourages low-frequency shape information and accurate color matching, and the SSIM encourages high-frequency details, for instance, the shading of the ears. The loss is

$$\mathcal{L}_{\text{face}} = \frac{\alpha}{N} \sum_{j \in \text{pixels}} \frac{1 - \text{SSIM}(x_j, \tilde{x}_j)}{2} + (1 - \alpha) |\mathbf{x}_j - \tilde{\mathbf{x}}_j| \quad (15)$$

where N is number of pixels times 3 color channels. Similarly to [53, 14], we use the blending coefficient of $\alpha = 0.85$. We optimize the loss using Adam [23], minibatch size 32, and initial learning rate 10^{-4} , dividing by 10 at iteration 30000 and 50000, for a total of 60000 iterations. We train on a single TITAN X Pascal GPU. $\frac{1}{4}$ -2 hours is sufficient for good results. Figure 6 shows the results of reoriented and relit faces from a held-out validation set. The input is on the left and the transformed outputs on the right. Top to bottom each row shows a different



Figure 7. Side-by-side output of DC-IGN [27] TOP and our facial transformer BOTTOM. We have grayscaled our image for a fairer comparison. Input on left, smoothly rotated faces on right. We emphasize here that the goal of DC-IGN is different to ours, since they learn unsupervised disentangling. We argue to use supervision when the information is accessible. Our use of supervision is evident in that we can quantitatively rotate our faces; whereas, DC-IGN cannot.

transformation, namely, lighting azimuth, lighting elevation, rotation azimuth, and rotation elevation. Faces inside the large green box span the transformation parameters seen at training time, those outside were not seen. We note the reconstruction fidelity and impressive ability to reorient out-of-plane rotations, but zooming in shows that the reconstructions lack high-frequency detail to be foolproof replicas of the input and the overall face shape changes slightly. For unseen transformation parameters, notice how faces just outside the green box are of similar quality to inside, but large deviations from the training set degrade. This is especially so for the geometric rotations, where the boundary surfaces (nose and chin in particular) begin to protrude from the face. Surprisingly, the shading of the faces is realistic outside of the box. We also compare against DC-IGN [27] in Figure 7. Our superior quality is partially down to better training, but also to improved alignment in feature-space, from supervised transformation information. Interpretability of our features allows for more accurate control over the azimuthal rotation. **Feature stability** In Figure 8 we test the feature stability under transformations of the input. We take an invariant representation of the data using L2-norms and relative phases, then measure the cosine similarity (top) and L2-distance (bottom) between a face and transformed versions of itself (blue), and we also compute these metrics between transformed versions of a face and a randomly selected face of another identity. There is a clear separation between faces of different identities for medium sized transformations, but this breaks down for large values of the parameters for geometric rotations. This is especially so, when the parameter values are close to the limit of the training data, as would be expected.

Real faces For fun, we feed images of real faces into our system, to recognize basic pose, shape, appearance, and lighting. We take internet images, cropping out background and hair. The system makes crude, but convincing enough matches to pose, skin tone, and lighting. The bottom image is particularly hard due to the side pose and lighting. This shows our system has learned a generalizable representation of faces, despite training

on artificial data.

4.3. Voxelized ShapeNets: 3D Transformations

For this experiment, we use the ModelNet10 subset of the ShapeNet dataset [52]. This consists of 3991 CAD models from 10 object categories. Specifically, we use the voxelized ModelNet10 provided by Maturana *et al.* [38], which is a volumetric binary occupancy grid of size 32x32x32.

The encoder-decoder architecture is similar to the variational auto-encoder architecture by Brock *et al.* [3], with the bottleneck of 200 units with equivariance to rotations about the y-axis. We also employ their variant of the binary cross entropy loss for training:

$$\mathcal{L}_{\text{bce}} = \sum_{i \in \text{voxels}} -\gamma t_i \log(o_i) - (1-\gamma)(1-t_i) \log(1-o_i), \quad (16)$$

where t_i are the target values rescaled to $[-1, 2]$, o_i is the output of the auto-encoder rescaled to $[0.1, 0.9999]$ and γ is set to 0.98 to compensate for the sparseness of volumetric data. We optimize the loss using Adam, minibatch size 16, and learning rate of 10^{-4} . See supplementary materials for details on classification.

5. Conclusion

We have presented a simple framework to learn deep feature-spaces, which disentangle both in-plane and out-of-plane transformations into an interpretable feature space, that also allows smooth interpolation. Our key innovation is the feature transform layer, which can be applied to both convolutional and fully-connected layers. The properties of the feature transform layer give our networks equivariance properties, that can help with generative and discriminative applications.

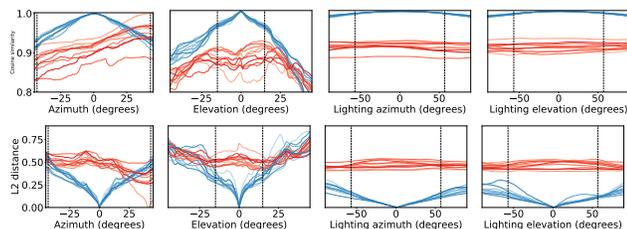


Figure 8. Pairs of images are compared to each other in feature space and similarity is measured using cosine similarity TOP and L2 distance BOTTOM. Pairs of images with same identity shown in blue, and pairs with different identities shown in red (10 each). Columns show left to right: sweeping of azimuth, elevation, lighting azimuth, and lighting elevation with all other parameters held. Dashed vertical lines show range of transformation values seen at training time. Ideally cosine similarity would be 1 everywhere for the blue lines, indicating perfect transformation invariance. For dissimilar faces, the red curves would be less than 1. We see that invariance to lighting is easy, even beyond the range of training examples (see central box in Figure 6). Elevation is particularly hard, so two features of the same person begin to differ at large elevations.

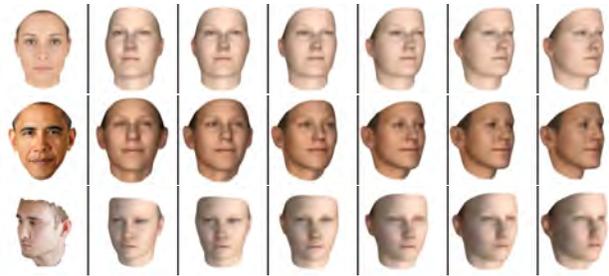


Figure 9. We pass images of real faces through our system re-orienting 50° from the initial pose, while fixing all other transformation parameters. Despite being trained on artificial data, the system is able to extract basic pose, shape, appearance and illumination. The system struggles to match shape properly, since these are far from the training set.

Limitations Our approach is supervised, so labeled examples are needed to span the space of transformations, preferably with little other variety in the images. Also, the feature space needs to be smooth, precluding mirroring.

Acknowledgements Support is from Fight for Sight UK, a Microsoft Research PhD Scholarship, NERC NE/P016677/1, and NERC NE/P019013/1.

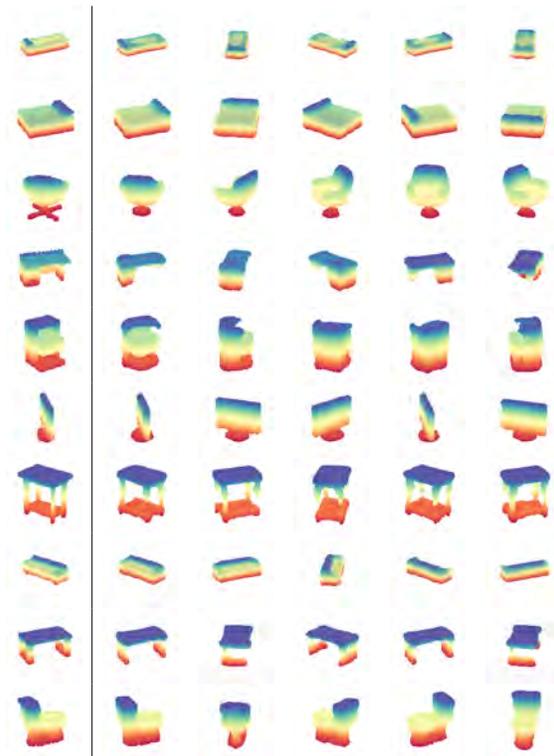


Figure 10. We pass randomly rotated volume of 10 categories from the test set (left) through our system re-orienting it by 0, 60, 120, 180, and 240 degrees from the initial pose. The system struggles to reconstruct thin shapes properly, which is a common problem due to sparseness of the volume occupancy.

References

- [1] P. F. Alcantarilla, A. Bartoli, and A. J. Davison. KAZE features. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI*, pages 214–227, 2012. 2
- [2] E. Barnard and D. Casasent. Invariance and neural nets. *IEEE Trans. Neural Networks*, 2(5):498–508, 1991. 3
- [3] A. Brock, T. Lim, J. Ritchie, and N. Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236*, 2016. 8
- [4] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013. 2
- [5] X. Chen, X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2172–2180, 2016. 3, 4, 5
- [6] T. Cohen and M. Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2990–2999, 2016. 2, 3, 6
- [7] T. S. Cohen and M. Welling. Transformation properties of learned visual representations. *CoRR*, abs/1412.7659, 2014. 3, 5
- [8] T. S. Cohen and M. Welling. Steerable CNNs. *CoRR*, abs/1612.08498, 2016. 3
- [9] J. L. Crowley and A. C. Parker. A representation for shape based on peaks and ridges in the difference of low-pass transform. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(2):156–170, 1984. 2
- [10] S. Dieleman, J. D. Fauw, and K. Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1889–1898, 2016. 2, 3
- [11] A. Dosovitskiy, J. Tobias Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1546, 2015. 3, 5
- [12] B. Fasel and D. Gatica-Perez. Rotation-invariant neoperceptron. In *18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China*, pages 336–339, 2006. 2, 3
- [13] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(9):891–906, 1991. 2
- [14] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016. 7
- [15] D. M. Gonzalez, M. Volpi, N. Komodakis, and D. Tuia. Rotation equivariant vector field networks. *CoRR*, abs/1612.09346, 2016. 2, 3, 6
- [16] D. M. Gonzalez, M. Volpi, and D. Tuia. Learning rotation invariant convolutional filters for texture classification. *CoRR*, abs/1604.06720, 2016. 2, 3
- [17] T. Hasegawa, M. Ambai, K. Ishikawa, G. Koutaki, Y. Yamauchi, T. Yamashita, and H. Fujiiyoshi. Multiple-hypothesis affine region estimation with anisotropic log filters. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 585–593, 2015. 2
- [18] J. F. Henriques and A. Vedaldi. Warped convolutions: Efficient invariance to spatial transformations. *CoRR*, abs/1609.04382, 2016. 3
- [19] G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer, 2011. 3, 4, 5
- [20] IEEE. *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009. 5, 6
- [21] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456, 2015. 5
- [22] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2017–2025, 2015. 3, 4, 5
- [23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5, 7
- [24] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 3
- [25] I. Kokkinos, M. Bronstein, and A. Yuille. Dense Scale Invariant Descriptors for Images and Surfaces. Research Report RR-7914, INRIA, Mar. 2012. 2
- [26] G. Koutaki and K. Uchimura. Scale-space processing using polynomial representations. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 2744–2751, 2014. 2
- [27] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2539–2547, 2015. 3, 5, 7
- [28] D. Laptev, N. Savinov, J. M. Buhmann, and M. Pollefeys. TI-POOLING: transformation-invariant pooling for feature learning in convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 289–297, 2016. 2, 3
- [29] H. Larochelle, D. Erhan, A. C. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pages 473–480, 2007. 5, 6
- [30] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems 2, [NIPS Conference, Denver, Colorado, USA, November 27-30, 1989]*, pages 396–404, 1989. 1, 3
- [31] K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 991–999, 2015. 1, 2

- [32] K. Lenc and A. Vedaldi. Learning covariant feature detectors. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, pages 100–117, 2016. 3, 5
- [33] R. Lenz. *Group Theoretical Methods in Image Processing*, volume 413 of *Lecture Notes in Computer Science*. Springer, 1990. 2
- [34] M. Lin, Q. Chen, and S. Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 3
- [35] T. Lindeberg. Generalized gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. *Journal of Mathematical Imaging and Vision*, 40(1):36–81, 2011. 2
- [36] R. Lippmann. An introduction to computing with neural nets. *IEEE Assp magazine*, 4(2):4–22, 1987. 1
- [37] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013. 5
- [38] D. Maturana and S. Scherer. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. In *IROS*, 2015. 8
- [39] R. Memisevic and G. E. Hinton. Learning to represent spatial transformations with factored higher-order Boltzmann machines. *Neural Computation*, 22(6):1473–1492, 2010. 3
- [40] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. <http://distill.pub/2016/deconv-checkerboard>. 7
- [41] E. Oyallon and S. Mallat. Deep roto-translation scattering for object classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2865–2873, 2015. 2, 3
- [42] P. Perona. Deformable kernels for early vision. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1991, 3-6 June, 1991, Lahaina, Maui, Hawaii, USA*, pages 222–227, 1991. 2
- [43] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3D structure from images. In *Advances In Neural Information Processing Systems*, pages 4997–5005, 2016. 3, 5
- [44] U. Schmidt and S. Roth. Learning rotation-aware features: From invariant priors to equivariant descriptors. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2050–2057, 2012. 6
- [45] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Trans. Information Theory*, 38(2):587–607, 1992. 2
- [46] K. Sohn and H. Lee. Learning invariant representations with local transformations. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012. 6
- [47] P. C. Teo. *Theory and Applications of Steerable Functions*. PhD thesis, Dept. Computer Science, Stanford University, 3 1998. 2
- [48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. 7
- [49] R. Wilson and H. Knutsson. Uncertainty and inference in the visual system. *IEEE Trans. Systems, Man, and Cybernetics*, 18(2):305–312, 1988. 2
- [50] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. *CoRR*, abs/1612.04642, 2016. 2, 3, 6
- [51] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1912–1920, 2015. 5
- [52] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1912–1920, 2015. 8
- [53] J. Xie, R. B. Girshick, and A. Farhadi. Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 842–857, 2016. 7
- [54] C. J. B. Yann LeCun, Corinna Cortes. The MNIST database of handwritten digits, 1999. 5
- [55] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 286–301, 2016. 3
- [56] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. Oriented response networks. *CoRR*, abs/1701.01833, 2017. 2, 3