VISUAL REPRESENTATIONS:

DEFINING PROPERTIES AND DEEP APPROXIMATIONS

Stefano Soatto
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095, USA
soatto@ucla.edu

Alessandro Chiuso
Dipartimento di Ingegneria dell'Informazione
Università di Padova
Via Gradenigo 6/b, 35131 Padova, Italy
alessandro.chiuso@unipd.it

ABSTRACT

Visual representations are defined in terms of minimal sufficient statistics of visual data, for a class of tasks, that are also invariant to nuisance variability. Minimal sufficiency guarantees that we can store a representation in lieu of raw data with smallest complexity and no performance loss on the task at hand. Invariance guarantees that the statistic is constant with respect to uninformative transformations of the data. We derive analytical expressions for such representations and show they are related to feature descriptors commonly used in computer vision, as well as to convolutional neural networks. This link highlights the assumptions and approximations tacitly assumed by these methods and explains empirical practices such as clamping, pooling and joint normalization.

1 INTRODUCTION

A *visual representation* is a function of visual data (images) that is "useful" to accomplish visual tasks. Visual tasks are decision or control actions concerning the surrounding environment, or *scene,* and its properties. Such properties can be geometric (shape, pose), photometric (reflectance), dynamic (motion) and semantic (identities, relations of "objects" within). In addition to such properties, the data also depends on a variety of *nuisance variables* that are irrelevant to the task. Depending on the task, they may include unknown characteristics of the sensor (intrinsic calibration), its inter-play with the scene (viewpoint, partial occlusion), and properties of the scene that are not directly of interest (*e.g.,* illumination).

We are interested in modeling and analyzing visual representations: How can we measure how "useful" one is? What guidelines or principles should inform its design? Is there such a thing as an *optimal* representation? If so, can it be computed? Approximated? Learned? We abstract (classes of) visual tasks to *questions about the scene*. They could be countable semantic queries (*e.g.,* concerning "objects" in the scene and their relations) or continuous control actions (*e.g.,* "in which direction to move next").

In Sect. 2.1 we formalize these questions using well-known concepts, and in Sect. 2.3 we derive an equivalent characterization that will be the starting point for designing, analyzing and learning representations.

1.1 RELATED WORK AND CONTRIBUTIONS

Much of Computer Vision is about computing functions of images that are "useful" for visual tasks. When restricted to subsets of images, *local descriptors* typically involve statistics of image gradients, computed at various scales and locations, pooled over spatial regions, variously normalized and quantized. This process is repeated hierarchically in a convolutional neural network (CNN), with weights inferred from data, leading to *representation learning* Ranzato et al. (2007); LeCun (2012); Simonyan et al. (2014); Serre et al. (2007); Bouvrie et al. (2009); Susskind et al. (2011); Bengio (2009). There are more methods than we can review here, and empirical comparisons (*e.g.,* Mikolajczyk et al. (2004)) have recently expanded to include CNNs. Unfortunately, many implementation details and parameters make it hard to draw general conclusions Chatfield et al. (2011). We take a different approach, and derive a formal expression for optimal representations from established principles of *sufficiency, minimality, invariance.* We show how existing descriptors are *related* to such representations, highlighting the (often tacit) underlying assumptions.

Our work relates most closely to Bruna & Mallat (2011); Anselmi et al. (2015) in its aim to construct and analyze representations for classification tasks. However, we wish to represent the *scene*, rather than the image, so occlusion and locality play a key role, as we discuss in Sect. 5. Also Morel & Yu (2011) characterize the invariants to certain nuisance transformations: Our models are more general, involving both the range and the domain of the data, although more restrictive than Tishby et al. (2000) and specific to visual data. We present an alternate interpretation of pooling, in the context of classical sampling theory, that differs from other analyses Gong et al. (2014); Boureau et al. (2010).

Our contributions are to (i) define minimal sufficient invariant representations and characterize them explicitly (Claim 1); (ii) show that local descriptors currently in use in Computer Vision can approximate such representations under very restrictive conditions (Claim 2 and Sec. 3.3); (iii) compute in closed form the minimal sufficient *contrast* invariant (14) and show how local descriptors relate to it (Rem. 2); show that such local descriptors can be implemented via linear convolutions and rectified linear units (Sect. 4.3); (iv) explain the practice of “joint normalization” (Rem. 5) and “clamping” (Sect. 3.3.1) as procedures to approximate the sufficient invariant; these practices are seldom explained and yet they have a significant impact on performance in empirical tests Kirchner (2015); (v) explain “spatial pooling” in terms of *anti-aliasing,* or local marginalization with respect to a small-dimensional nuisance group, in convolutional architectures (Sec. 4.1-4.2). In the Appendix we show that an ideal representation, if generatively trained, maximizes the information content of the representation (App. A).

2 Characterization and properties of representations

Because of uncertainty in the mechanisms that generate images, we treat them as realizations of random vectors x (past/training set) and y (future/test set), of high but finite dimension. The scene they portray is infinitely more complex.[1] Even if we restrict it to be a static “model” or “parameter” θ, it is in general infinite-dimensional. Nevertheless, we can *ask questions* about it. The number of questions (“classes”) K is large but finite, corresponding to a partition of the space of scenes, represented by samples $\{\theta_1, \dots, \theta_K\}$. A simple model of image formation including scaling and occlusion Dong & Soatto (2014) is known as the Lambert-Ambient, or LA, model.

2.1 Desiderata

Among (measurable) functions of past data (statistics), we seek those useful for a class of tasks. Abstracting the task to *questions* about the scene, “useful” can be measured by uncertainty reduction on the answers, captured by the *mutual information* between the data x and the object of interest θ. While a representation can be no more informative than the data,[2] ideally it should be no less, *i.e.,* a *sufficient statistic*. It should also be “simpler” than the data itself, ideally *minimal*. It should also discount the effects of nuisance variables $g \in G$, and ideally be *invariant* to them. We use a superscript to denote a collection of t data points (the history up to t, if ordered), $x^t \doteq \{x_1, \dots, x_t\}$.

> Thus, *a representation is any function ϕ constructed using past data x^t that is useful to answer questions about the* scene *θ given future data y it generates, regardless of nuisance factors g.*

[1] Scenes are made of surfaces supporting reflectance functions that interact with illumination, etc. No matter how many images we already have, even a single static scene can generate infinitely many different ones.

[2] Data Processing Inequality, page 88 of Shao (1998).

An optimal representation is *a minimal sufficient statistic for a task that is invariant to nuisance factors*. In Sec. 2.3 we introduce the SA Likelihood, an optimal representation, and in subsequent sections show how it can be approximated.

2.2 Background

The data X is a random variable with samples x, y; the model θ is unknown in the experiment $E = \{x, \theta, p_\theta(x)\}$ where $p_\theta(x)$ is the probability density function of X, that depends on the parameter θ, evaluated at the sample x; a *statistic* T is a function of the sample; it is *sufficient* (of x for θ) if $X \mid T = \tau$ does not depend on θ;[3] it is *minimal* if it is a function of all other sufficient statistics[4]. If θ is treated as a random variable and a prior is available, ϕ is Bayesian sufficient[5] if $p(\theta|\phi(x^t)) = p(\theta|x^t)$.

If T is minimal, any smaller[6] U entails "information loss." If θ was a discrete random variable, the information content of T could be measured by uncertainty reduction: $\mathbb{H}(\theta) - \mathbb{H}(\theta|T(X))$, which is the mutual information[7] between θ and T and $\mathbb{H}$ denotes entropy Cover & Thomas (1991); furthermore, $T(X) \in \arg\inf_\phi \mathbb{H}(\theta|\phi(X))$, where the infimum is with respect to measurable functions and is in general not unique.

Consider a set G of transformations g of the data x, $g(x)$, which we denote simply as gx. A function $\phi_G(x)$ is G-invariant if $\phi_G(gx) = \phi_G(x)$ for all $g \in G$. The *sensitivity* of ϕ to G is $S = \|\frac{\partial \phi(gx)}{\partial g}\|$ where ϕ is assumed to be differentiable. By definition, an invariant has zero sensitivity to G.

The *Likelihood function* is $L(\theta; x) \doteq p_\theta(x)$, understood as a function of θ for a fixed sample x, sometimes written as $p(x|\theta)$ even though θ is not a random variable. Theorem 3.2 of Pawitan (2001) can be extended to an infinite-dimensional parameter θ (Theorem 6.1 of Bahadur (1954)):

Theorem 1 (The likelihood function as a statistic)**.** *The likelihood function $L(\theta; x)$ is a minimal sufficient statistic of x for θ.*

2.3 Nuisance management: Profile, marginal, and SAL likelihoods

A *nuisance* $g \in G$ is an unknown "factor" (a random variable, parameter, or transformation) that is not of interest and yet it affects the data. Given $p_\theta(\cdot)$, when g is treated as a parameter that transforms the data via $g(y) \doteq gy$, then $p_{\theta,g}(y) \doteq p_\theta(gy)$; when it is treated as a random variable, $p_\theta(y|g) \doteq p_\theta(gy)$. The *profile likelihood*

$$p_{\theta,G}(y) \doteq \sup_{g \in G} p_{\theta,g}(y) \tag{1}$$

where the nuisance has been "maxed-out" is G-invariant. The *marginal likelihood*

$$p_\theta(y|G) \doteq \int_G p_\theta(y|g) dP(g) \tag{2}$$

is invariant *only if* $dP(g) = d\mu(g)$ is the constant[8] measure on G. Both are *sufficient invariants*, in the sense that they are invariant to G *and* are minimal sufficient. This counters the common belief that "invariance trades off selectivity." In Rem. 1 we argue that both can be achieved, at the price of complexity.

Computing the profile likelihood in practice requires reducing G to a countable set $\{g_1, \dots, g_N\}$ of *samples*,[9] usually at a loss. The tradeoff is a subject of sampling theory, where samples can

[3] Definition 3.1 of Pawitan (2001) or Sec. 6.7 of DeGroot (1989), page 356

[4] If U is sufficient, then the sigma algebra $\sigma(T) \subset \sigma(U)$, DeGroot (1989), page 368.

[5] The two are equivalent for discrete random variables, but pathological cases can arise in infinite dimensions Blackwell & Ramamoorthi (1982).

[6] In the sense of inclusion of sigma algebras, $\sigma(U) \subset \sigma(T)$.

[7] See Cover & Thomas (1991) eq. 2.28, page 18.

[8] Base, or Haar measure if G is a group. It can be improper if G is not compact.

[9] Note that N can be infinite if G is not compact.

be generated *regularly*, independent of the signal being sampled, or *adaptively*.[10] In either case, the occurrence of spurious extrema ("aliases") can be mitigated by retaining *not* the value of the function at the samples, $p_{\theta,g_i}(y)$, but an *anti-aliased* version consisting of a weighted average around the samples:

$$\hat{p}_{\theta,g_i}(y) \doteq \int p_{\theta,g_i}(gy)w(g)d\mu(g) \tag{3}$$

for suitable weights w.[11] When the prior $dP(g) = w(g)d\mu(g)$ is positive and normalized, the previous equation (anti-aliasing) can be interpreted as *local marginalization*, and is often referred to as *mean-pooling*. The approximation of the profile likelihood obtained by sampling and anti-aliasing, is called the *SA (sampled anti-aliased) likelihood, or SAL:*

$$\hat{p}_{\theta,G}(y) = \max_i \hat{p}_{\theta,g_i}(y) = \max_i \int p_{\theta,g_i}(gy)dP(g). \tag{4}$$

The maximization over the samples in the above equation is often referred to as *max-pooling*.

Claim 1 (The SAL is an optimal representation)**.** *Let the joint likelihood $p_{\theta,g}$ be smooth with respect to the base measure on G. For any approximation error ϵ, there exists an integer $N = N(\epsilon)$ number of samples $\{g_i\}_{i=1}^N$ and a suitable (regular or adaptive) sampling mechanism so that the SAL $\max_i \hat{p}_{\theta,g_i}$ approximates to within ϵ the profile likelihood $\sup_{g\in G} p_{\theta,g}$, after normalization, in the sense of distributions.*

For the case of (conditionally) Gaussian models under the action of the translation group, the claim follows from classical sampling arguments. More generally, an optimal representation is difficult to compute. In the next section, we show a first example when this can be done.

Remark 1 (Invariance, sensitivity and "selectivity")**.** *It is commonly believed that invariance comes at the cost of discriminative power. This is partly due to the use of the term invariance (or "approximate invariance" or "stability") to denote* sensitivity, *and of the term "selectivity" to denote maximal invariance. A function is* insensitive *to g if small changes in g produce small changes in its value. It is invariant to g if it is constant as a function of g. It is a maximal invariant if equal value implies equivalence up to G (Sect. 4.2 of Shao (1998), page 213). It is common to decrease sensitivity with respect to a transformation g by* averaging *the function with respect to g, a lossy operation in general. For instance, if the function is the image itself, it can be made insensitive to rotation about its center by averaging rotated versions of it. The result is an image consisting of concentric circles, with much of the informative content of the original image gone, in the sense that it is not possible to reconstruct the original image. Nevertheless, while invertibility is relevant for reconstruction tasks, it is not necessarily relevant for classification, so it is possible for the averaging operation to yield a sufficient invariant, albeit not a maximal one.*

Thus, *one can have invariance while retaining sufficiency*, albeit generally not maximality: The profile likelihood, or the marginal likelihood with respect to the uniform measure, are sufficient statistics and are (strictly) invariant. The price to have both is *complexity*, as both are infinite-dimensional in general. However, they can be approximated, which is done by sampling in the SA likelihood.

2.4 A FIRST EXAMPLE: LOCAL REPRESENTATIONS/DESCRIPTORS

Let the task be to decide whether a (future) image y is of a scene θ given a *single* training image x of it, which we can then assume to be the scene itself: $x = \theta$. Nuisances affecting y are limited

[10] $\{g_i\}_{i=1}^N$ are generated by a (deterministic or stochastic) mechanism ψ that depends on the data and respects the structure of G. If G is a group, this is known as a *co-variant detector*: It is a function ψ that (is Morse in g, *i.e.,* it) has isolated extrema $\{g_i(y)\}_{i=1}^N = \{g \mid \nabla_G\psi(y,g) = 0\}$ that equivary: $\nabla_G\psi(\tilde{g}y, g_i) = 0 \Rightarrow \nabla_G\psi(y, \tilde{g}g_i) = 0$ for all i and $\tilde{g} \in G$. The samples $\{g_i\}_{i=1}^N$ define a reference frame in which an invariant can be easily constructed in a process known as *canonization* Soatto (2009): $\phi(y) \doteq \{g_i^{-1}(y)y \mid \nabla_G\psi(y,g_i) = 0\}$.

[11] For regular sampling of stationary signals on the real line, optimal weights for reconstruction can be derived explicitly in terms of spectral characteristics of the signal, as done in classical (Shannon) sampling theory. More in general, the computation of optimal weights for function-valued signals defined on a general group G for tasks other than reconstruction, is an open problem. Recent results Chen & Edelsbrunner (2011) show that diffusion on the location-scale group *typically* reduce the incidence of topological features such as aliases in the filtered signal. Thus, low-pass filtering such as (generalized) Gaussian smoothing as done here, can have anti-aliasing effects.

to translation parallel to the image plane, scaling, and changes in pixel values that do not alter relative order.[12] Under these (admittedly restrictive) conditions, the SAL can be easily computed and corresponds to known "descriptors." Note that, by definition, x and y must be generated by the same scene θ for them to "correspond."

SIFT Lowe (2004) performs canonization[10] of local similarity transformations via adaptive sampling of the planar translation-scale group (extrema of the difference-of-Gaussians operator in space and scale), and planar rotation (extrema of the magnitude of the oriented gradient). Alternatively, locations, scales and rotation can be sampled regularly, as in "dense SIFT." Regardless, on the domain determined by each sample, g_i, SIFT computes a weighted histogram of gradient orientations. Spatial regularization anti-aliases translation; histogram regularization anti-aliases orientation; scale anti-aliasing, however, is not performed, an omission corrected in DSP-SIFT Dong & Soatto (2015).

In formulas, if $\alpha(y) = \angle\nabla y \in \mathbb{S}^1$ is a direction, $\theta = x_i$ is the image restricted to a region determined by the reference frame g_i, centered at $(u_i, v_i) \in \mathbb{R}^2$ axis-aligned and with size $s_i > 0$, we have[13]

$$\phi_{x_i}(y) = \int \kappa_\sigma(u_i - \tilde{u}, v_i - \tilde{v})\kappa_\epsilon(\angle\nabla x(\tilde{u},\tilde{v}), \alpha(y))\|\nabla x(\tilde{u},\tilde{v})\|\mathcal{E}_{s_i}(\sigma)d\tilde{u}d\tilde{v}d\sigma \tag{5}$$

and $\phi_{\text{sift}}(\alpha) = [\phi_{x_{11}}(y), \dots, \phi_{x_{44}}(y)]$ is a sampling on a 4×4 grid, with each sample assumed independent and $\alpha = \angle\nabla y$ quantized into 8 levels. Variants of SIFT such as HOG differ in the number and location of the samples, the regions where the histograms are accumulated and normalized. Here κ_σ and κ_ϵ are Parzen kernels (bilinear in SIFT; Gaussian in DSP-SIFT) with parameter $\sigma, \epsilon > 0$ and $\mathcal{E}_s$ is an exponential prior on scales. Additional properties of SIFT and its variants are discussed in Sect. 3.3, as a consequence of which the above approximates the SAL for translation-scale and contrast transformation groups.

Claim 2 (DSP-SIFT)**.** *The continuous extension of DSP-SIFT Dong & Soatto (2015) (5) is an anti-aliased sample of the profile likelihood (4) for $G = SE(2)\times\mathbb{R}^+\times\mathcal{H}$ the group of planar similarities transformations and local contrast transformations, when the underlying scene $\theta = x_i$ has locally stationary and ergodic radiance, and the noise is assumed Gaussian IID with variance proportional to the norm of the gradient.*

The proof follows from a characterization of the maximal invariant to contrast transformations described in Sect. 3.3.

Out-of-plane rotations induce a scene-shape-dependent deformation of the domain of the image that cannot be determined from a single training image, as we discuss in Sect. 3.

When interpreting local descriptors as samples of the SAL, they are usually assumed *independent*, an assumption lifted in Sect. 4 in the context of convolutional architectures.

3 A more realistic instantiation

Relative motion between a non-planar scene and the viewer generally triggers occlusion phenomena. These call for the representation to be *local*. Intrinsic variability of a non-static scene must also be taken into account in the representation. In this section we describe the approximation of the SAL under more realistic assumptions than those implied by local, single-view descriptors such as SIFT.

3.1 Occlusion, clutter and "receptive fields"

The data y has many components, $y = \{y_1, \dots, y_{M_y}\}$, only an unknown subset of which from the scene or object of interest θ (the "visible" set $V \subset D = \{1, \dots, M_y\}$). The rest come from clutter,

[12] The planar translation-scale group can be taken as a very crude approximation of the transformation induced on the image plane by spatial translation. Contrast transformations (monotonic continuous transformations of the range of the image) can be interpreted as crude approximations of changes of illumination.

[13] Here $gy(u_i, v_i) = y(u_i + u, v_i + v)$ for the translation group and $gy(u_i, v_i) = y\,(\sigma u_i + u, \sigma v_i + v))$ for translation-scale. In general, $gy - x \neq y - g^{-1}x$, unless $gy - x = 0$. If $p_x(y(u)) \doteq q(y(u) - x(u))$ for some q is a density function for the random variable $y(u)$, in general $q(gy(u) - x(u)) \neq q(y(u) - g^{-1}x(u))$, unless the process y is G-stationary independent and identically distributed (IID), in which case $p_x(gy) = p_{g^{-1}x}(y)$. Note that the marginal density of the gradient of natural images is to a reasonable approximation invariant to the translation-scale group Huang & Mumford (1999).

occlusion and other phenomena unrelated to θ, although they may be informative as "context." We indicate the restriction of y to V as $y_{|_V} = \{y_j,\ j \in V\}$. Since the visible set is not known, profiling $p_{\theta,G}(y) = \max_{i,V\in\mathcal{P}(D)} p_{\theta,g_i}(y_{|_V})$ requires searching over the power set $\mathcal{P}(D)$. To make computation tractable, we can restrict the visible set V to be the union of "receptive fields" V_j, that can be obtained by transforming[14] a "base region" $\mathcal{B}_0$ ("unit ball," for instance a square patch of pixels with an arbitrary "base size," say 10×10) with group elements $g_j \in G$, $V_j \doteq g_j\mathcal{B}_0$: $V = \bigcup_{j=1}^M g_j\mathcal{B}_0$ where the number of receptive fields $M \ll M_y$. Thus V is determined by the reference frames (group elements) $\{g_j\}_{j=1}^M$ of receptive fields that are "active," which are unknown a-priori.

Alternatively, we can marginalize V by computing, for every class (represented by a hidden variable θ_k as discussed next) and every receptive field (determined by g_j as above), conditional densities $p_\theta(y|\theta_k, g_j)$ that can be averaged with respect to weights w_{jk}, trained to *select* (if sparse along j) and *combine* (by averaging along k) local templates via

$$\sum_{j,k} p_\theta(y|\theta_k, g_j)w_{jk} \tag{6}$$

where the weights or "filters" w_{jk}, if positive and normalized,[15] are interpreted as probabilities $w_{jk} = p_\theta(\theta_k, g_j)$. To make this marginalization tractable, we write the first term as

$$p_\theta(y_{|_{V_j}}, y_{|_{V_j^c}}|\theta_k, g_j) = p(y_{|_{V_j}}|\theta_k, g_j)p_\theta(y_{|_{V_j^c}}) \propto p(y_{|_{V_j}}|\theta_k, g_j) \tag{7}$$

where we have assumed that the second factor is constant, thus ignoring *photometric context* beyond the receptive fields, *e.g.,* due to mutual illumination. Under these assumptions, we have

$$p_{\theta,g_i}(y) = \sum_{j,k} p(y_{|_{V_j}}|\theta_k, g_ig_j)p_{\theta,g_i}(g_j\theta_k) \tag{8}$$

where the first term in the sum is known as a *"feature map"* and we have assumed that both $g_i, g_j \in G$ for simplicity, with $g_{ij} = g_ig_j$. The order of operations (deformation by g_i and selection by g_j) is arbitrary, so we assume that the selection by g_j is applied first, and then the nuisance-induced deformation g_i, so $p_{\theta,g_i}(g_jy) \propto p_{\theta,e}(g_{ij}y)$, where e is the identity of the group G.

3.2 INTRA-CLASS VARIABILITY AND DEFORMABLE TEMPLATES

For category-level recognition, the parameter space can be divided into K classes, allowing variability of θ within each class. Endowing the parameter space with a distribution $p(\theta|k)$ requires defining a probability density in the space of shapes, reflectance functions etc. Alternatively one can capture the variability θ induces on the data. For any scene θ_k from class k, one can consider a single image generated by it $x_k \sim p_{\theta_k}(x)$ as a "template" from which any other datum from the same class can be obtained by the (transitive) action of a group $g_k \in G$. Thus if $y \sim p_{\theta_k}(y)$ with $\theta_k \sim p(\theta|k)$, which we indicate with $y \sim p(y|k)$, then we assume that there exists a g_k such that $y = g_k^{-1}x_k$, so

$$\begin{aligned} p(y|k) &= \int p(y|x_k, g_k)dP(x_k, g_k|\theta_k)dP(\theta_k|k) \\ &= \int p(g_ky|\theta_k)dP(g_k|\theta_k) \end{aligned} \tag{9}$$

where we have used the fact that $p(y|x_k, g_k, k) = \delta(y - g_k^{-1}x_k)$ and that only one sample of θ_k is given, so all variability is represented by g_k and x_k conditionally on θ_k.[16] For this approach to work, g_k has to be sufficiently complex to allow x_k to "reach" every datum y generated by an element[17] of

[14] The action of a group g on a set $\mathcal{B} \subset D$ is defined as $g\mathcal{B} \subset D$ such that $g(y_{|_\mathcal{B}}) = y_{|_{g\mathcal{B}}}$.

[15] Note that current convolutional architectures rectify and normalize the feature maps, not the filters. However, learned biases, as well as rectification and normalization at each layer, may partly compensate for it.

[16] In the last expression we assumed that x_k and g_k are conditionally independent given θ_k, *i.e.,* that the image formation process (noise) and deformation are independent once the scene θ_k is given.

[17] The group has to act transitively on x_k. For instance, in Grenander (1993) g_k was chosen to belong to the entire (infinite-dimensional) group of domain diffeomorphisms.

the class. Fortunately, the density on a complex group can be reduced to a joint density on G^M, the mutual configuration of the receptive fields, as we will show.

The restriction of g_k to the domain of the receptive field $V_j = g_j\mathcal{B}_0$ is indicated by $\{g_{kj}\}_{j=1}^M$, defined by $g_{k_j}x = g_k x \; \forall \; x \in V_j$. Then, we can consider the global group nuisance g_i, the selector of receptive fields g_j and the local restriction of the intra-class group g_k, assumed $(d(g_k, e))$ small, as all belonging to the same group G, for instance affine or similarity transformations of the domain and range space of the data.

Starting from (8), neglecting g_i for the moment, we have[18]

$$p_\theta(y) = \sum_{j,k} p(y_{|V_j}|\theta_k, g_j)p_\theta(\theta_k, g_j) \tag{10}$$

$$= \sum_{j,k} \int_{G^M} p(y_{|V_j}|\theta_k, g_{k_j})dP(\{g_{k_j}\}|\theta_k)p_\theta(\theta_k, g_j) \tag{11}$$

and bringing back the global nuisance g_i,

$$p_{\theta,G}(y) = \max_i \underbrace{\sum_{j,k} \int_{G^M} p(g_{ik_j}y_{|V_j}|\theta_k)dP_G(\{g_{k_j}\}|\theta_k)p_\theta(\theta_k, g_j)}_{p_\theta(g_i y)} \tag{12}$$

where the measure in the last equation is made invariant to $g_i \in G$. The feature maps $p(g_i g_{k_j} y_{|V_j}|\theta_k)$ represent the photometric component of the likelihood. The geometric component is the relative configuration of receptive fields $\{g_{k_j}\}$, which is class-dependent but G-invariant. The inner integral corresponds to "mean pooling" and the maximization to "max pooling." The sum over j, k marginalizes the local classes θ_k, or *"parts"* and selects them to compose the hypothesis θ.

To summarize, g_i are the samples of the nuisance group in (4); g_j are the local reference frames that define each receptive field in (8); g_k are the global deformations that define the variability induced by a class k on a template in (9). The latter are in general far more complex than the former, but their restriction to each receptive field, g_{k_j}, can be approximated by an affine or similarity transformation and hence composed with g_i and g_j.

Note that (11) can be interpreted as a model of a three-layer neural network: The visible layer, where y lives, a hidden layer, where the feature maps $p(y_{|V_j}|\theta_k, g_{k_j})$ live, and an output layer that, after rectification and normalization, yields an approximation of the likelihood $p_\theta(y)$. Invariance to G can be obtained via a fourth layer outputting $p_{\theta,G}(y)$ by max-pooling third-layer outputs $p_\theta(g_i y)$ for different g_i in (12).

3.3 Contrast invariance

Contrast is a monotonic continuous transformation of the (range space of the) data, which can be used to model changes due to illumination. It is well-known that the curvature of the level sets of the image is a maximal invariant Alvarez et al. (1993). Since it is everywhere orthogonal to the level sets, the gradient orientation is also a maximal contrast invariant. Here we compute a contrast invariant by marginalizing the norm of the gradient of the test image (thus retaining its orientation) in the likelihood function of a training image. Since the action of contrast transformations is spatially independent, in the absence of other nuisances we assume that the gradient of the test image y can be thought of as a noisy version of the gradient of the training image x, *i.e.,*

$$\nabla y \sim \mathcal{N}(\nabla x, \epsilon^2) \tag{13}$$

and compute the density of y given x marginalized with respect to contrast transformations $\mathcal{H}$ of y.

Theorem 2 (Contrast-invariant sufficient statistic)**.** *The likelihood of a training image x at a given pixel, given a test image y, marginalized with respect to contrast transformations of the latter, is*

[18] Here we condition on the restrictions g_{k_j} of g_k on the receptive fields V_j so that, by definition, $p(y_{|V_j}|\theta_k, g_j, g_k) = p(y_{|V_j}|\theta_k, g_{k_j})$.

given by

$$\boxed{p_x(y|\mathcal{H}) \doteq p(\angle\nabla y|\nabla x) = \frac{1}{\sqrt{2\pi\epsilon^2}} \exp\left(-\frac{1}{2\epsilon^2}\sin^2(\angle\nabla y - \angle\nabla x)\|\nabla x\|^2\right) M} \tag{14}$$

where, if we call $\Psi(a) \doteq \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-\frac{1}{2}\tau^2}\,d\tau$ *for any* $a \in \mathbb{R}$, *and* $m \doteq \cos(\angle\nabla y - \angle\nabla x)\|\nabla x\|$, *then*

$$M = \frac{\epsilon e^{-\frac{(m)^2}{2\epsilon^2}}}{\sqrt{2\pi}} + m - m\Psi\left(-\frac{m}{\epsilon}\right). \tag{15}$$

The expression in (14) *is, therefore, a minimal sufficient statistic of* y *that is invariant to contrast transformations.*

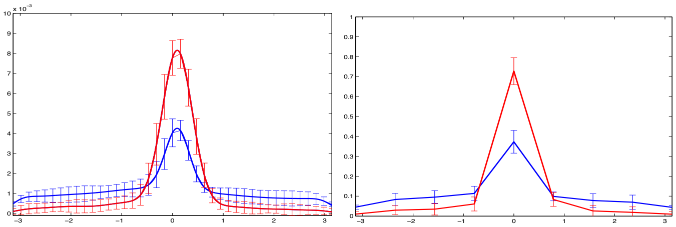

Figure 1: **SIFT integrand** (18) **(red) vs. marginalized likelihood** (14) **(blue)** *computed for a random patch on* $\alpha \in [-\pi, \pi]$ *(left), and on a regular sub-sampling of* 8 *orientations (right). Several random tests are shown as mean and error-bars corresponding to three standard deviations across trials.*

Remark 2 (Relation to SIFT)**.** *Compared to* (14), *SIFT (i) neglects the normalization factor* $\frac{M}{\sqrt{2\pi}\epsilon}$, *(ii) replaces the kernel*

$$\tilde{\kappa}_\epsilon(\alpha) \doteq \exp\left(\frac{1}{2\epsilon^2}\sin^2(\alpha)\right) \simeq \exp\left(\frac{1}{2\epsilon^2}\alpha^2\right) \tag{16}$$

with a bilinear *one* κ_ϵ *defined by*

$$\kappa_\epsilon(\alpha) \doteq \begin{cases} \frac{\alpha+\epsilon}{\epsilon^2} & \alpha \in [-\epsilon,\ 0] \\ \frac{\epsilon-\alpha}{\epsilon^2} & \alpha \in [0,\ -\epsilon] \end{cases} \tag{17}$$

and, finally, (iii) multiplies the result by the norm of the gradient, obtaining the sift integrand

$$\phi_{\text{sift}}(\angle\nabla y|\nabla x) = \kappa_\epsilon(\angle\nabla y - \angle\nabla x)\|\nabla x\| \tag{18}$$

To see this, calling $\alpha = \angle\nabla y - \angle\nabla x$ *and* $\beta = \|\nabla x\| > 0$, *notice that*

$$\kappa_\epsilon(\alpha)\beta = \kappa_{\epsilon\beta}(\alpha\beta) \simeq \exp\left(\frac{1}{2\epsilon^2\beta^2}\alpha^2\beta^2\right) \simeq \tilde{\kappa}_\epsilon(\alpha) \tag{19}$$

where the left-hand side is (18) *and the right-hand side is* (14) *or* (23).

We make no claim that this approximation is good, it just happens to be the choice made by SIFT, and the above just highlights the relation to the contrast invariant (14).

Remark 3 (Uninformative training images)**.** *It is possible that* $\frac{m}{\epsilon} \simeq 0$ *holds uniformly (in* α*) provided* $\frac{\gamma}{\epsilon} \ll 1$, i.e., , *if the modulus of the gradient* ∇x *is very small as compared to the standard deviation* ϵ. *Under such circumstances, the training image* x *is essentially constant ("flat"), and the conditional density* $p(\alpha|\nabla x)$ *becomes uniform*

$$\begin{array}{rcl} p(\alpha|\nabla x) & \simeq & \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-\frac{1}{2\epsilon^2}\gamma^2\left(1-\langle\overline{\nabla y},\overline{\nabla x}\rangle^2\right)} \frac{\epsilon}{\sqrt{2\pi}} \\ & = & \frac{1}{\sqrt{2\pi\epsilon^2}} \frac{\epsilon}{\sqrt{2\pi}} = \frac{1}{2\pi} \end{array} \tag{20}$$

where the approximation holds given that $e^{-\frac{1}{2\epsilon^2}\gamma^2\left(1-\langle\overline{\nabla y},\overline{\nabla x}\rangle^2\right)} \simeq 1$ *when* $\frac{\gamma}{\epsilon} \ll 1$. *This is unlike SIFT* (18), *that becomes zero when the norm of the gradient goes to zero.*

Note that, other than for the gradient, the computations in (14) can be performed point-wise, so for an image or patch with pixel values y_i, if $\alpha_i(y) \doteq \angle \nabla y_i$, we can write

$$p(\alpha|\nabla x) = \prod_i p(\alpha_i|\nabla x_i). \tag{21}$$

We often omit reference to contrast transformations $\mathcal{H}$ in $p_x(y|\mathcal{H})$, when the argument α makes it clear we are referring to a contrast invariant. The width of the kernel ϵ is a design (regularization) parameter.

Remark 4 (Invariance for x)**.** *Note that* (21) *is invariant to contrast transformations of y, but* not *of x. This should not be surprising, since high-contrast training patches should yield tests with high confidence, unlike low-contrast patches. However, when the training set contains instances that are subject to contrast changes, such variability must be managed.*

To eliminate the dependency on $\|\nabla x\|$, consider a model where the noise is proportional to the norm of the gradient:

$$\nabla y \sim \mathcal{N}\left(\nabla x, \tilde{\epsilon}^2\right) \tag{22}$$

where $\tilde{\epsilon}(\|\nabla x\|) = \epsilon\|\nabla x\|$. Under this noise model, the sufficient contrast invariant (14) *becomes*

$$p_x(y|\mathcal{H}) \doteq p(\angle \nabla y|\nabla x) = \frac{1}{\sqrt{2\pi\epsilon^2}} \exp\left(-\frac{1}{2\epsilon^2}\sin^2(\angle\nabla y - \angle\nabla x)\right) M \tag{23}$$

and M has the same expression (15) *but with $m = \cos(\angle\nabla y - \angle\nabla x)$. Thus a simple approach to managing contrast variability of x* in addition to *y is to use the above expression in lieu of* (14).

Remark 5 (Joint normalization)**.** *If we consider only* affine *contrast transformations $ax + b$ where a, b are assumed to be constant on a patch V which contains all the cells C_i where the descriptors are computed*[19] *it is clear that to recapture invariance w.r.t. the scale factor a it is necessary and sufficient that $p(\angle\nabla y|\nabla x(v_i)) = p(\angle\nabla y|a\nabla x(v_i))$, $\forall v_i \in V$. We shall now illustrate how this invariance can be achieved.*

Assume that data generating model (13) *is replaced by the distribution-dependent model*

$$\nabla y \sim \mathcal{N}(\nabla x, \epsilon^2(p_x)) \quad \epsilon^2(p_x) = \sigma^2 \mathbb{E}_x\|\nabla x\|^2 = \sigma^2 \int \|\nabla x\|^2 p_x(\nabla x) d\nabla x \tag{24}$$

where the noise variance ϵ^2 depends linearly on the average squared gradient norm (w.r.t. the distribution $p_x(\nabla x)$); σ^2 is fixed constant. The resulting marginal distribution for the gradient orientation becomes

$$\bar{p}_x(y|\mathcal{H}) \doteq \bar{p}(\angle\nabla y|\nabla x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\sin^2(\angle\nabla y - \angle\nabla x)\frac{\|\nabla x\|^2}{\mathbb{E}_x\|\nabla x\|^2}\right)\bar{M} \tag{25}$$

where, defining $\bar{m} \doteq \cos(\angle\nabla y - \angle\nabla x)\frac{\|\nabla x\|}{\sqrt{\mathbb{E}_x\|\nabla x\|^2}}$,

$$\bar{M} = \frac{\sigma e^{-\frac{(\bar{m})^2}{2\sigma^2}}}{\sqrt{2\pi}} + \bar{m} - \bar{m}\Psi\left(-\frac{\bar{m}}{\sigma}\right). \tag{26}$$

Equation (25) *is clearly invariant to affine transformations of the image values $x(v) \to ax(v) + b$, $\forall v \in V$.*[20] *It is a trivial calculation to show that using $\bar{p}(\angle\nabla y|\rho)$ in lieu of $p(\angle\nabla y|\rho)$, the result is invariant w.r.t affine transformations.*

To obtain a sampled version of this normalization the expected squared gradient norm can be replaced with the sample average on the training patch

$$\hat{\rho}^2 \doteq \frac{1}{N_{pix}} \sum_{i=1}^{N_{pix}} \|\nabla x(v_i)\|^2$$

[19] SIFT divides each patch into a 4×4 grid of cells C_i, $i = 1, .., 16$

[20] Note that an affine transformation on the image values $x(v) \to ax(v) + b$, $\forall v \in V$, induces a scale transformation on the distribution $p_x(\nabla x)$ so that $p_{ax+b}(\rho) = \frac{1}{a}p_x(\rho/a)$ and therefore the average squared gradient is scaled by a^2, i.e. $\mathbb{E}_{ax+b}\|\rho\|^2 = a^2\mathbb{E}_x\|\rho\|^2$.

so that (24) *becomes,*

$$\nabla y \sim \mathcal{N}(\nabla x, \epsilon^2(V)) \quad \epsilon^2(V) = \sigma^2 \hat{\rho}^2 = \sigma^2 \frac{1}{N_{pix}} \sum_{i=1}^{N_{pix}} \|\nabla x(v_i)\|^2 \tag{27}$$

where $v_i \in V$, $i = 1, .., N_{pix}$ *are the pixel locations in the training patch* V*. This procedure is known as* "joint normalization"*, and is simply equivalent to normalizing the patch in pre-processing by dividing by the average gradient norm.*

3.3.1 Clamping Gradient Orientation Histograms

Local descriptors such as SIFT commonly apply a *"clamping"* procedure to modify a (discretized, spatially-pooled, un-normalized) density ϕ_{sift} of the form (18), by clipping it at a certain threshold τ, and re-normalizing:

$$\phi_{\text{clamp}}(\alpha|x) = \frac{\min\{\phi_{\text{sift}}(\alpha|x), \tau\}}{\int_{\mathbb{S}^1} \min\{\phi_{\text{sift}}(\alpha|x), \tau\} d\alpha} \tag{28}$$

where $\alpha = \angle \nabla y \in \mathbb{S}^1$ is typically discretized into 8 bins and τ is chosen as a percentage of the maximum, for instance $\tau = 0.2 * \max_\alpha \phi_{\text{sift}}(\alpha|x)$. Although clamping has a dramatic effect on performance, with high sensitivity to the choice of threshold, it is seldom explained, other than as a procedure to "reduce the influence of large gradient magnitudes" Lowe (2004).

Here, we show empirically that (18) becomes closer to (14) after clamping for certain choices of threshold τ and sufficiently coarse binning. Fig. 2 shows that, without clamping, (18) is considerably more peaked than (14) and has thinner tails. After clamping, however, the approximation improves, and for coarse binning and threshold between around 20% and 30% the two are very similar.

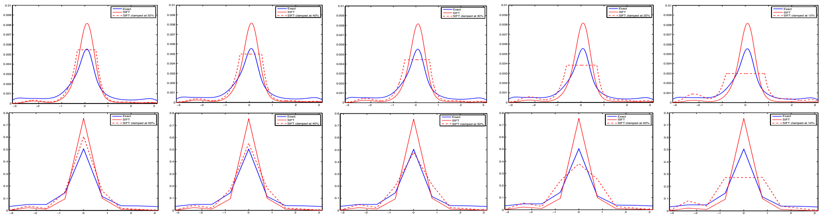

Figure 2: **Clamping effects:** *For* $\alpha \in [-\pi, \pi]$ *(abscissa), the top shows the marginalized likelihood* $p(\alpha|\nabla x)$ (14) *(blue), the SIFT integrand* (18) *(solid red), and its clamped version* (28) *(dashed red) for thresholds ranging from* 50% *to* 10% *of its maximum. The bottom shows the same discretized to* 8 *orientation bins. The clamping approximation is sensible only for coarse binning, and heavily influenced by the choice of threshold. For an* 8*-bin histogram, the best approximation is typically achieved with clamping threshold between* 10% *and* 30% *of the maximum; note that* Lowe *(*2004*) empirically chose* 20%.

3.4 Rotation Invariance

Canonization Soatto (2009) is particularly well suited to deal with planar rotation, since it is possible to design co-variant detectors with few isolated extrema. An example is the local maximum of the norm of the gradient along the direction $\alpha = \hat{\alpha}_l(x)$. Invariance to $G = SO(2)$ can be achieved by retaining the samples $\{p_\theta(\alpha|\hat{\alpha}_l)\}_{l=1}^L$. When no consistent (sometimes referred to as "stable") reference $\hat{\alpha}_l$ can be found, it means that there is no co-variant functional with isolated extrema with respect to the rotation group, which means that the data is already invariant to rotation.

Note that, again, planar rotations can affect both the training image x and the test image y. In some cases, a consistent reference (canonical element) is available. For instance, for geo-referenced scenes $L = 1$, and the projection of the gravity vector onto the image plane , $\hat{\alpha}$, provides a canonical reference unless the two are orthogonal:

$$p_\theta(\alpha|G) = p_\theta(\alpha|\hat{\alpha}). \tag{29}$$

4 Deep Convolutional Architectures

In this section we study the approximation of (12) implemented by convolutional architectures. Starting from (12) for a particular class and a finite number of receptive fields we notice that, since the "true scene" θ and the nuisances g are unknown, we cannot factor the likelihood $p_{\theta,g}(y)$ into a product of p_{θ,g_j}, which would correspond to a "bag-of-words" discarding the dependencies in $dP_G(\{g_j\}|\theta)$. Convolutional architectures (CNNs) promise to capture such dependencies by hierarchical decomposition into progressively larger receptive fields. Each "layer" is a collection of separator (hidden) variables (nodes) that make lower layers (approximately) conditionally independent.

4.1 Stacking simplifies nuisance marginalization

We show this in several steps. We first argue that managing the group of diffeomorphisms can be accomplished by independently managing *small* diffeomorphisms in each layer. We use marginalization, but a similar argument can be constructed for max-out or the SA likelihood. Then, we leverage on the local restrictions induced by receptive fields, to deal with occlusion, and argue that such small diffeomorphisms can be reduced locally to a *simpler group* (affine, similarity, location-scale or even translations, the most common choice in convolutional architectures). Then global marginalization of diffeomorphisms can be accomplished by local marginalization of the reduced group in each layer. The following lemma establishes that global diffeomorphisms can be approximated by the composition of small diffeomorphisms.

Lemma 1. *Let $g \in G$, $g : D \to D$ be an orientation-preserving diffeomorphism of a compact subset D of the real plane, $e \in G$ the identity, and $d(e, g)$ the "size" of g. Then for any $\epsilon > 0$ there exists an $N < \infty$ and $g^1, \dots, g^N$ such that $g = g^1 \circ g^2 \cdots \circ g^N$ and $d(e, g^i) < \epsilon \ \forall\, i = 1, \dots, N$.*

Now for two layers, let $g = g^1 \circ g^2$, with $g^1, g^2 \sim p(g)$ drawn independently from a prior on G. Then $p(g|g^1, g^2) = \delta(g - g^1 \circ g^2)$ (or a non-degenerate distribution if g^i are approximated by elements of the reduced group). Then let $\theta^1 = g^2\theta$, or more generally let θ^1 be defined such that $\theta^1 \perp g^1 \mid \theta, g^2$. We then have

$$\begin{aligned} p_G(y|\theta) &= \int p(y|\theta, g)dP(g) = \int p(y|\theta^1, g^1, g)dP(\theta^1, g^1, g^2|\theta, g)dP(g) = \quad (30)\\ &= \int p(y|\theta^1, g^1)dP(g^1|\theta)p(\theta^1|\theta, g^2)dP(g^2|\theta)d\theta^1 \quad (31)\end{aligned}$$

where we have also used the fact that $y \perp \theta \mid \theta^1$. Once the separator variable θ^1 is reduced to a number K_1 of filters, we have

$$p_G(y|\theta) \simeq \sum_{k=1}^{K_1} \int p(y|\theta^1_k, g^1)dP(g^1|\theta) \int p(\theta^1_k|\theta, g^2)dP(g^2|\theta) \simeq \sum_{k=1}^{K_1} p_G(y|\theta^1_k)p_G(\theta^1_k|\theta) \quad (32)$$

in either case, by extending this construction to $L = N$ layers, we can see that marginalization of each layer can be performed with respect to (conditionally) independent diffeomorphisms that can be chosen to be small per the Lemma above.

Claim 3. *Marginalization of the likelihood with respect to an arbitrary diffeomorphism $g \in G$ can be achieved by introducing layers of hidden variables θ^l $l = 1, \dots, L$ and independently marginalizing* small *diffeomorphisms $g^l \in G$ at each layer.*

The next step is to restrict the marginalization to each receptive field, at which point it can be approximated by a reduced subgroup, or the (linear) generators.

4.2 Hierarchical decomposition of the likelihood

Let

$$p_G(y|\theta) \doteq \int_G p(y|\theta, g)dP(g) \quad (33)$$

be the marginal likelihood with respect to some prior on G and introduce a layer of "separator variables" θ^1 and group actions g, defined such that $y \perp \theta \mid (\theta^1, g^1)$. This can always be done by choosing $\theta^1 = y$; we will address non-trivial choices later. In either case, forgoing the subscript G,

$$p(y|\theta) = \int p(y|\theta^1, g^1)dP(\theta^1, g^1|\theta). \quad (34)$$

If θ^1 and g^1 take a finite number K_1 and L_1 of values $\{\theta^1_1, \dots, \theta^1_{K_1}\}$ (filters) and $\{g^1_1, \dots, g^1_{L_1}\}$, then the above reduces to a sum over $k = 1, \dots, K_1$ and $\ell_1 = 1, \dots L_1$; the conditional likelihoods $\{p(y|\theta^1_1, g^1_j), \dots, p(y|\theta^1_{K_1}, g^1_j)\}$ are the *feature maps*. If y has dimensions $N \times M$ and the group actions g^1_j are taken to be pixel wise translations across the image plane, so that $L_1 = N \times M$, the feature maps $p(y|\theta^1, g^1)$ can be represented as a tensor with dimensions $N \times M \times K_1$. One can repeat the procedure for new separator variables that take K_2 possible values, and group actions g^2 that take $L_2 = N_1 \times M_1$ values; the filters θ^2 *must be supported on the entire feature maps* $p(y|\theta^1, g^1)$ (*i.e.,* take values in $N_1 \times M_1 \times K_1$) for the sum over $k = 1, \dots, K_1$ to implement the marginalization above

$$p(y|\theta) = \sum_{\ell_2=1}^{L_2} \sum_{j=1}^{K_2} \left[\sum_{\ell_1=1}^{L_1} \sum_{k=1}^{K_1} p(y|\theta^1_k, g^1_{\ell_1}) p(\theta^1_k, g^1_{\ell_1}|\theta^2_j, g^2_{\ell_2}) \right] p(\theta^2_j, g^2_{\ell_2}|\theta). \tag{35}$$

The sum is implemented in convolutional networks by the use of translation invariant filters:

1. At the first layer the support of θ^1 is a small fraction of $N \times M$ and g^1 acts on y so that[21] $p(y|\theta^1_k, g^1_{\ell_1}) = p(g^1_{\ell_1} y|\theta^1_k) = p(y_{|V_j}|\theta^1_k)$.
2. At the second layer the filter $p(\theta^1_k, g^1_{\ell_1}|\theta^2_j, g^2_{\ell_2})$ is nonzero for a finite (and small) number of group actions $g^1_{\ell_1}$ and also satisfies the shift invariant (convolutional) property $p(\theta^1_k, g^1_{\ell_1}|\theta^2_j, g^2_{\ell_2}) = p(\theta^1_k, g^2_{\ell_2} g^1_{\ell_1}|\theta^2_j)$

The third dimension of the filters is the number of feature maps in the previous layer.

4.3 Approximation of the first layer

Each node in the first layer computes a local representation (5) using parent node(s) as a "scene." This relates to existing convolutional architectures where nodes compute the response of *a rectified linear unit (ReLu) to a filter bank*. For simplicity we restrict G to the translation group, thus reducing (5) to SIFT, but the arguments apply to similarities.

A ReLu response at (u, v) to an oriented filter bank $\mathcal{G}$ with scale σ and orientation α is given by $R_+(\alpha, u, v, \sigma) = \max(0, \mathcal{G}(u, v; \sigma, \alpha) * x(u, v))$. Let $\mathcal{N}(u, v; \sigma)$ be a Gaussian, centered in (u, v) with isotropic variance σI, $\nabla \mathcal{N}(u, v; \sigma) = [\frac{\partial \mathcal{N}}{\partial u}(u, v; \sigma) \ \frac{\partial \mathcal{N}}{\partial v}(u, v; \sigma)]$, $r(\alpha) = [\cos\alpha \ \sin\alpha]^T$. Then $\mathcal{G}(u, v; \sigma, \alpha) \doteq \nabla \mathcal{N}(u, v; \sigma) r(\alpha)$ is a directional filter with principal orientation α and scale σ. Omitting rectification for now, the response of an image to a filter bank obtained by varying $\alpha \in [-\pi, \ \pi]$, at each location (u, v) and for all scales σ is obtained as

$$\begin{aligned} R(\alpha, u, v, \sigma) &= \mathcal{G}(u, v; \sigma, \alpha) * x(u, v) = \mathcal{N}(u, v; \sigma) * \nabla x(u, v) r(\alpha) && (36) \\ &= \mathcal{N}(u, v; \sigma) * \left\langle \frac{\nabla x(u, v)}{\|\nabla x(u, v)\|}, r(\alpha) \right\rangle \|\nabla x(u, v)\| && (37) \\ &= \int \mathcal{N}(u - \tilde{u}, v - \tilde{v}; \sigma) \kappa(\angle \nabla x(\tilde{u}, \tilde{v}), \alpha) \|\nabla x(\tilde{u}, \tilde{v})\| d\tilde{u} d\tilde{v} && (38) \end{aligned}$$

where κ, the cosine function, has to be rectified for the above to approximate a histogram, $\kappa_+(\alpha) = \max(0, \cos\alpha)$ which yields SIFT. Unfortunately, in general the latter does not equal $\max(0, \mathcal{G} * x)$ for one cannot simply move the maximum inside the integral. However, under conditions on x, which are typically satisfied by natural images, this is the case.

Claim 4. *Let $\mathcal{G}$ be positive, smooth and have a small effective support $\sigma < \infty$. I.e., $\forall\, \epsilon_1, \epsilon_2 \, \exists\, \sigma \,|\, \mathrm{vol}(\mathcal{G}(\tilde{u}, \tilde{v}; \sigma, \alpha) \geq \epsilon_1) < \epsilon_2$. Let x have a sparse and continuous gradient field, so that for every α the domain of x can be partitioned in three (multiply-connected) regions $D_+(\alpha)$, $D_-(\alpha)$ and the remainder (the complement of their union), where the projection of the gradient in the direction α is, respectively, positive, negative, and negligible, and $d(\alpha) > 0$ the minimum distance*

[21] There is a non-trivial approximation here, namely that context is neglected when assuming that the likelihood $p(g^1_{\ell_1} y|\theta^1_k)$ depends only on the restriction of y to the receptive field V_j; see also Section 3.1 and equation (7).

between the regions D_+ and D_-. Then, provided that $\sigma \leq \min_\alpha d(\alpha)$, we have that

$$\underbrace{\max(0, \mathcal{G}(u, v; \sigma, \alpha) * x(u, v))}_{\text{ReLu}} \simeq \underbrace{\int \mathcal{N}(u - \tilde{u}, v - \tilde{v}; \sigma)\kappa_+(\angle \nabla x(\tilde{u}, \tilde{v}), \alpha)\|\nabla x(\tilde{u}, \tilde{v})\| d\tilde{u} d\tilde{v}}_{\text{sift}} \tag{39}$$

4.4 Stacking information

A local hierarchical architecture allows approximating the SA likelihood $p_{\theta,G}(\cdot)$ by reducing nuisance management to local marginalization and max-out of simple group transformations. The SA likelihood $p_{\theta,G}(y)$ is an optimal representation for any query on θ given data y. For instance, for classification, the representation $p_{\theta,G}(y)$ is itself the classifier (discriminant). Thus, if we could compute an optimal classifier, the representation would be the identity; vice-versa, if we could compute the optimal representation, the classifier would be a threshold. In practice, one restricts the family of classifiers – for instance to soft-max, or SVM, or linear – leaving the job of feeding the most informative statistic to the classifier. In a hierarchical architecture, this is the feature maps in the last layer. This is equivalent to neglecting the global dependency $p_{\theta,G}(y|\theta^L)$ on θ at the last layer. The information loss inherent in this choice is the loss of assuming that θ^L are independent (whereas they are only independent conditioned on θ).

An optimal representation with restricted complexity $L < \infty$, therefore, maximizes the independence of the components of θ^L, or equivalently the independence of the components of y given θ^L. Using those results, one can show that the information content of a representation ((47) in App. A) grows with the number of layers L.

5 Discussion

For the likelihood interpretation of a CNN put forward here to make sense, training should be performed *generatively*, so fixing the class θ_k one could sample (hallucinate) future images y from the class. However neither the architecture not the training of current CNN incorporate mechanisms to enforce the statistics of natural images.

In this paper we emphasize the role of the *task* in the representation: If nothing is known about the task, nothing interesting can be said about the representation, and the only optimal one is the trivial one that stores all the data. This is because the task could end up being a query about the value of a particular pixel in a particular image. Nevertheless, there may be many different tasks that share the same representation by virtue of the fact that they share the same nuisances. In fact, the task affects what are nuisance factors, and the nuisance factors affect the design and learning of the representation. For some complex tasks, writing the likelihood function may be prohibitively complex, but some classes of nuisances such as changes of illumination or occlusions, are common to many tasks.

Note that, by definition, a nuisance is not informative. Certain transformations that are nuisances for some tasks may be informative for others (and therefore would not be called nuisances). For instance, viewpoint is a nuisance in object detection tasks, as we want to detect objects regardless of where they are. It is, of course, not a nuisance for navigation, as we must control our pose relative to the surrounding environment. In some cases, nuisance and intrinsic variability can be entangled, as for the case of intra-class deformations and viewpoint-induced deformations in object categorization. Nevertheless, the deformation would be informative *if it was known or measured*, but since it is not, it must be marginalized.

Our framework does not require nuisances to have the structure of a group. For instance, occlusions do not. Invariance and sensitivity are still defined, as a statistic is invariant if it is constant with respect to variations of the nuisance. What is not defined is the notion of *maximal invariance*, that requires the orbit structure. However, in our theory maximal invariance is not the focus. Instead, *sufficient invariance* is.

The literature on the topic of representation is vast and growing. We focus on visual representations, where several have been active. Anselmi et al. (2015) have developed a theory of representation aiming at approximating maximal invariants, which restricts nuisances to (locally) compact groups and therefore do not explicitly handle occlusions. Both frameworks achieve invariance at the expense of discriminative power, whereas in our framework both can be attained at the cost of complexity. Patel et al. (2015), that appeared after earlier drafts of this manuscript were made public, instead of of starting from principles and showing that they lead to a particular kind of computational architecture, instead assume a particular architecture and interpret it probabilistically, similarly to Ver Steeg & Galstyan (2015) that uses total correlation as a proxy of information, which is related to our App. A. However, there the representation is defined absent a task, so the analysis does not account for the role of nuisance factors.

In particular, Anselmi et al. (2015) define a $\mathcal{G}$-invariant representation μ of a (single) image I as being *selective* if $\mu(I) = \mu(I') \Rightarrow I \sim I'$ for all I, I', *i.e.,* if it is a *maximal invariant*. But while equivalence to the data up to the action of the nuisance group is critical for *reconstruction*, it is not necessary for other tasks. Furthermore, for non-group nuisances, such as occlusions, a maximal invariant cannot be constructed. Instead, given a task, we replace maximality with *sufficiency* for the task, and define at the outset an optimal representation to be a *minimal sufficient invariant statistic,* or "sufficient invariant," approximated by the *SAL Likelihood.* The construction in Anselmi et al. (2015) guarantees maximality for compact groups. Similarly, Sundaramoorthi et al. (2009) have shown that maximal invariants can be constructed even for diffeomorphisms, which are infinite-dimensional and non-compact. In practice, however, *occlusions and scaling/quantization* break the group structure, and therefore a different approach is needed that relies on sufficiency, not maximality, as we proposed here. To relate our approach to Anselmi et al. (2015), we note that the orbit probability of Def. 4.2 is given by

$$\rho_I(A) = P(\{g \mid gI \in A\}) \tag{40}$$

and is used to induce a probability on I, via $P(I)[A] = P(\{g \mid gI \in A\})$. On the other hand, we define the minimal sufficient invariant statistic as the marginalized likelihood

$$p_{\theta,G}(y) \doteq \int p_{\theta,g}(y) dP(g) \tag{41}$$

where y is a (future) image, and θ is the *scene*. If we consider the scene to be comprised of a set of images $A = \theta$, and the future image $y = I$, then we see that the OP is a marginalized likelihood where $dP(g) = d\mu(g)$ is the Haar measure, and $p_{\theta,g}(y) = \delta(gy \cap \theta)$. Thus, substitutions $G \leftarrow \mathcal{G}$, $\theta \leftarrow A$, $y \leftarrow I$ yield

$$P(I)[A] = p_{\theta,G}(y) \tag{42}$$

for the particular choice of Haar measure and impulsive density $p_{\theta,g}$. The TP representation can also be understood as a marginalized likelihood, as $\Psi(I)[A]$ is the $\mathcal{G}$-marginalized likelihood of I given A when using the uniform prior and an impulsive conditional $p_{A,g}(I)$:

$$\Psi(I)[A] = \int_{\mathcal{G}} p(gI|A) d\mu(g). \tag{43}$$

Finally, our treatment of representations is not biologically motivated, in the sense that we simply define optimal representations from first principles, without regards for whether they are implementable with biological hardware. However, we have established connections with both local descriptors and deep neural networks, that were derived using biological inspiration.

Acknowledgments

Work supported by FA9550-15-1-0229, ARO W911NF-15-1-0564, ONR N00014-15-1-2261, and MIUR RBFR12M3AC "Learning meets time." We are appreciative of discussions with Tomaso Poggio, Lorenzo Rosasco, Stephane Mallat, and Andrea Censi.

References

Alvarez, L., Guichard, F., Lions, P. L., and Morel, J. M. Axioms and fundamental equations of image processing. *Arch. Rational Mechanics*, 123, 1993. 7

Anselmi, F., Rosasco, L., and Poggio, T. On invariance and selectivity in representation learning. *arXiv preprint arXiv:1503.05938*, 2015. 2, 13, 14

Bahadur, R. R. Sufficiency and statistical decision functions. *Annals of Mathematical Statistics*, 25 (3):423–462, 1954. 3

Bengio, Y. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2 (1):1–127, 2009. 2

Blackwell, D. and Ramamoorthi, R.V. A bayes but not classically sufficient statistic. *The Annals of Statistics*, 10(3):1025–1026, 1982. 3

Boureau, Y.-L., Ponce, J., and LeCun, Y. A theoretical analysis of feature pooling in visual recognition. In *Proc. of Intl Conf. on Mach. Learning*, pp. 111–118, 2010. 2

Bouvrie, J. V., Rosasco, L., and Poggio, T. On invariance in hierarchical models. In *NIPS*, pp. 162–170, 2009. 2

Bruna, J. and Mallat, S. Classification with scattering operators. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, 2011. 2

Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. The devil is in the details: an evaluation of recent feature encoding methods. 2011. 2

Chen, C. and Edelsbrunner, H. Diffusion runs low on persistence fast. In *ICCV*, pp. 423–430, 2011. 4

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley, 1991. 3

Creutzig, F., Globerson, A., and Tishby, N. Past-future information bottleneck in dynamical systems. *Phys. Rev. Lett. E*, 79(4):19251–19255, 2009. 17

DeGroot, M. H. *Probability and statistics*. Addison-Wesley, 1989. 3

Dong, J. and Soatto, S. *Machine Learning for Computer Vision*, chapter Visual Correspondence, the Lambert-Ambient Shape Space and the Systematic Design of Feature Descriptors. R. Cipolla, S. Battiato, G.-M. Farinella (Eds), Springer Verlag, 2014. 2

Dong, J. and Soatto, S. Domain size pooling in local descriptors: Dsp-sift. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, 2015. 5

Fraser, D. and Naderi, A. Minimal sufficient statistics emerge from the observed likelihood function. *International Journal of Statistical Science*, 6:55–61, 2007. 17

Gong, Y., Wang, L., Guo, R., and Lazebnik, S. Multi-scale orderless pooling of deep convolutional activation features. *arXiv preprint arXiv:1403.1840*, 2014. 2

Grenander, U. *General Pattern Theory*. Oxford University Press, 1993. 6

Hinkley, D. Predictive likelihood. *The Annals of Statistics*, pp. 718–728, 1979. 19

Huang, J. and Mumford, D. Statistics of natural images and models. In *Proc. CVPR*, pp. 541–547, 1999. 5

Kirchner, M. R. Automatic thresholding of SIFT descriptors. *Technical Report*, 2015. 2

LeCun, Y. Learning invariant feature hierarchies. In *ECCV*, pp. 496–505, 2012. 2

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *IJCV*, 2(60):91–110, 2004. 5, 10

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Gool, L. Van. A comparison of affine region detectors. *IJCV*, 1(60):63–86, 2004. 2

Morel, J. M. and Yu, G. Is sift scale invariant? *Inverse Problems and Imaging*, 5(1):115–136, 2011. 2

Patel, A., Nguyen, T., and Baraniuk, R. A probabilistic theory of deep learning. *arXiv preprint arXiv:1504.00641*, 2015. 14

Pawitan, Y. *In all likelihood: Statistical modeling and inference using likelihood*. Oxford, 2001. 3, 19

Ranzato, M., Huang, F. J., Boureau, Y.-L., and LeCun, Y. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR*, pp. 1–8, 2007. 2

Serre, T., Oliva, A., and Poggio, T. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429, 2007. 2

Shao, J. *Mathematical Statistics*. Springer Verlag, 1998. 2, 4

Simonyan, K., Vedaldi, A., and Zisserman, A. Learning local feature descriptors using convex optimisation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2(4), 2014. 2

Soatto, S. Actionable information in vision. In *Proc. of the Intl. Conf. on Comp. Vision*, October 2009. 4, 10, 17, 18

Sundaramoorthi, G., Petersen, P., Varadarajan, V. S., and Soatto, S. On the set of images modulo viewpoint and contrast changes. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, June 2009. 14

Susskind, J., Memisevic, R., Hinton, G. E., and Pollefeys, M. Modeling the joint density of two images under a variety of transformations. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, pp. 2793–2800, 2011. 2

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Proc. of the Allerton Conf.*, 2000. 2

Ver Steeg, G. and Galstyan, A. Maximally informative hierarchical representations of high-dimensional data. *in depth*, 13:14, 2015. 14

A Quantifying the information content of a representation

In general we do not know the likelihood, but we have a collection of *samples* $x_t \sim p_{\theta,g_t}(x)$, each generated with some nuisance g_t, which we can use to infer, or "learn" an approximation of the SAL likelihood Fraser & Naderi (2007).

$$\phi_\theta(\cdot) \doteq p_\theta(\cdot) \quad \simeq \quad \hat{p}_{X^t}(\cdot),\ X^t \sim p_\theta(\cdot) \qquad \text{(empirical likelihood)} \tag{44}$$

$$\phi_{\theta,G}(\cdot) \doteq p_{\theta,G}(\cdot) \quad \simeq \quad \hat{p}_{X^t,G}(\cdot) \qquad \text{(profile likelihood)} \tag{45}$$

$$\phi_{\theta,G}(y,x^t) \doteq \phi_{\theta,G}(y)\phi_\theta(x^t) \quad \simeq \quad \hat{p}_{X^t,G}(y)\hat{p}_{X^t}(x^t) \propto \hat{p}_{x^t,G}(y) \ \text{(learned representation)} \tag{46}$$

Depending on modeling choices made, including the number of samples N, the sampling mechanism, and the priors for local marginalization, the resulting representation will be "lossy" compared to the data. Next we quantify such a loss.

A.1 Informative content of a representation

The scene θ is in general infinite-dimensional. Thus, the informative content of the data cannot be directly quantified by mutual information $\mathbb{I}(\theta; x^t)$. In fact, $\mathbb{H}(\theta) = \infty$ and therefore, no matter how many (finite) data x^t we have, $\mathbb{I}(\theta; x^t) = \mathbb{H}(\theta) - \mathbb{H}(\theta|x^t) = \infty - \infty$ is not defined. Similarly, $\mathbb{I}(\theta; \phi(x^t))$ is undefined and therefore mutual information cannot be used directly to measure the informative content of a representation, or to infer the most informative statistic.

The notion of *Actionable Information* Soatto (2009) as $\mathcal{H}(x^t) \doteq \mathbb{H}(\phi_{\theta,G}(x^t))$ where $\phi_{\theta,G}$ is a maximal G-invariant, can be used to bypass the computation of $\mathbb{H}(\theta)$: Writing formally

$$\mathbb{I}(\theta; \phi_{\theta,G}(y)) = \mathbb{H}(\phi_{\theta,G}(y)) - \mathbb{H}(\phi_{\theta,G}(y)|\theta) = \mathcal{H}(x^t) - \mathbb{H}(\phi_{\theta,G}(y)|\theta) \tag{47}$$

we see that, if we were given the scene θ, we could generate the data y, under the action of some nuisance g, up to the residual modeling uncertainty, which is assumed white and zero-mean (lest the mean and correlations of the residual can be included in the model). Similarly, we can generate a maximal invariant $\phi_{\theta,G}(y)$ up to a residual with constant entropy; therefore, the statistic which maximizes $\mathbb{I}(\theta; \phi_{\theta,G}(y))$ is the one that maximizes the first term in (47), $\mathbb{H}(\phi_{\theta,G}(y))$. This formal argument allows defining the most informative statistic as the one that maximizes Actionable Information, $\hat{\phi}_{\theta,G} = \arg\max_{\phi_{\theta,G}} \mathbb{H}(\phi_{\theta,G}(x^t)) \doteq \mathcal{H}(x^t)$, bypassing the computation of the entropy of the "scene" θ. Note that the task still influences the information content of a representation, by defining what are the nuisances G, which in turn affect the computation of actionable information.

An alternative approach is to measure the informative content of a representation bypassing consideration of the scene is described next.

Definition 1 (Informative Content of a Representation)**.** *The information a statistic ϕ of x^t conveys on θ is the information it conveys on a task T (*e.g.*, a question on the scene θ), regardless of nuisances $g \in G$:*

$$\mathbb{I}(gT; \phi(x^t)) = \mathbb{H}(gT) - \mathbb{H}(gT|\phi(x^t)) \quad \forall\, g \in G \tag{48}$$

If the task is *reconstruction* (or prediction) $T = y$, where past data x^t and future data y are generated by the same scene θ, then the definition above relates to the past-future mutual information Creutzig et al. (2009) except for the role of the nuisance g. The following claim shows that an ideal representation, as previously defined in terms of minimal sufficient invariant statistic, maximizes information.

Claim 5. *Let past data x^t and future data y, used to accomplish a task T, be generated by the same scene θ. Then the representation $\phi_{\theta,G}$ maximizes the informative content of a representation.*

The next claim relates a representation to Actionable Information.

Claim 6. *If ϕ maximizes Actionable Information, it also maximizes the informative content of a representation.*

Note that maximizing Actionable Information is a stronger condition than maximizing the information content of a representation. Since Actionable Information concerns maximal invariance, sufficiency is automatically implied, and the only role the task plays is the definition of the nuisance group G. This is limiting, since we want to handle nuisances that do not have the structure of a group, and therefore Definition 1 affords more flexibility than Soatto (2009).

The next two claims characterize the maximal properties of the profile likelihood. We first recall that the marginalized likelihood is invariant *only if* marginalization is done with respect to the base (Haar) measure, and in general it is not a maximal invariant, as one can show easily with a counter-example (*e.g.,* a uniform density). On the other hand, the profile likelihood is by construction invariant regardless of the distribution on G, but is also – in general – not maximal. However, it is maximal under general conditions on the likelihood. To see this, consider $p_\theta(\cdot)$ to be given; for a free variable x, it can be written as a map q: $p_\theta(x) \doteq q(\theta, x)$. For a fixed x, q is a function of θ. If q is constant along x (the level curves are straight lines), then in general $q(\theta, y) = q(\theta, x)$ for all θ does *not* imply $y = x$. Indeed, y can be arbitrarily different from x. Even if q is non-degenerate (non-constant along certain directions), but presents *symmetries*, it is possible for different $y \neq x$ to yield the same $q(\theta, y) = q(\theta, x)$ for all θ. However, under *generic conditions* $q(\theta, y) = q(\theta, x)$ for all θ implies $y = x$. Now consider $p_{\theta,G}(\cdot)$ to be given; for a free variable x the map can be written using a function q such that $p_{\theta,G}(x) = \min_g q(\theta, gx)$. Note that $p_{\theta,G}$ is, by construction, invariant to G. Also note that, following the same argument as above, the invariant is not maximal, for it is possible for $p_{\theta,G}(x) = p_{\theta,G}(y)$ for all θ and yet x and y are not equivalent (equal up to a constant g: $y = gx$). However, if the function q is *generic*, then the invariant is maximal. In fact, let $\hat{g}(\theta, x) = \arg\min q(\theta, gx)$, so that $p_{\theta,G}(x) = q(\theta, \hat{g}(\theta, x)x)$. If we now have $q(\theta, \hat{g}(\theta, y)y) = q(\theta, \hat{g}(\theta, x)x)$ for all θ, then we can conclude, based on the argument above, that $\hat{g}(\theta, x)x = \hat{g}(\theta, y)y$. Since x and y are fixed, and the equality is for all θ, we can conclude that $\hat{g}(\theta, y)^{-1}\hat{g}(\theta, x)$ is independent of θ. That allows us to conclude the following.

Claim 7 (Maximality of the profile likelihood)**.** *If the density $p_\theta(x)$ is generic with respect to x, then $p_{\theta,G}(\cdot)$ is a maximal G-invariant.*

Since we do not have control on the function $p_{\theta,G}(\cdot)$, which is instead in general constructed from data, it is legitimate to ask what happens when the generic condition above is not satisfied. Fortunately, distributions that yield non-maximal invariants can be ruled out as uninformative at the outset:

Claim 8 (Non-maximality and non-informativeness)**.** *If q is such that, for any $x \neq y$ we have $q(\theta, \hat{g}(\theta, x)x) = q(\theta, \hat{g}(\theta, y)y)$ for all θ, then $q_{\theta,G}(\cdot)$ is uninformative.*

This follows from the definition of information, for any statistic T.

As we have pointed out before, what matters is not that the invariant be *maximal*, but that it be *sufficient*. As anticipated in Rem. 1, we can achieve invariance with *no sacrifice* of discriminative power, albeit at the cost of complexity.

B PROOFS

Theorem 1

Proof. Pick any θ_0, define $T(x) \doteq \frac{L(\cdot;x)}{L(\theta_0;x)}$ and $f(T(x), \theta) \doteq \frac{L(\theta;x)}{L(\theta_0;x)}$ and apply the factorization lemma with $h(x) = L(\theta_0; x)$. □

Theorem 2

Proof. We denote with $\overline{\nabla y} \doteq \frac{\nabla y}{\|\nabla y\|}$ the normalized gradient of y, and similarly for x; Φ maps it to polar coordinates $(\alpha, \rho) = \Phi(\nabla y)$ and $(\beta, \gamma) = \Phi(\nabla x)$, where

$$\alpha \doteq \angle \nabla y \quad \rho \doteq \|\nabla y\| \quad \beta \doteq \angle \nabla x \quad \gamma \doteq \|\nabla x\|.$$

The conditional density of ∇y given ∇x takes the polar form

$$\begin{array}{rcl} p(\rho, \alpha|\nabla x) & = & p(\nabla y|\nabla x)_{\nabla y = \Phi^{-1}(\rho,\alpha)}\rho \\ & & p(\nabla y|\nabla x) = \frac{1}{2\pi\epsilon^2} e^{-\frac{1}{2\epsilon^2}\|\nabla y - \nabla x\|^2}. \end{array} \tag{49}$$

Defining $(\nabla x)_i$ to be the i-th component of ∇x, (49) can be expanded as

$$p(\rho, \alpha|\nabla x) = \rho \frac{1}{2\pi\epsilon^2} e^{-\frac{1}{2\epsilon^2}\left[(\rho\cos(\alpha)-(\nabla x)_1)^2+(\rho\sin(\alpha)-(\nabla x)_2)^2\right]} \tag{50}$$

and the exponent is

$$\begin{aligned}(\rho\cos(\alpha)-(\nabla x)_1)^2+(\rho\sin(\alpha)-(\nabla x)_2)^2 &= \rho^2 - 2\rho((\nabla x)_1\cos\alpha + (\nabla x)_2\sin\alpha) + \|\nabla x\|^2 \\ &= \left(\rho - \gamma\langle\overline{\nabla y}, \overline{\nabla x}\rangle\right)^2 + \gamma^2\left(1 - \langle\overline{\nabla y}, \overline{\nabla x}\rangle^2\right)\end{aligned} \tag{51}$$

We are now interested in the marginal of (50) with respect to ρ, *i.e.,*

$$p(\alpha|\nabla x) = \int_0^\infty p(\rho, \alpha|\nabla x)\, d\rho. \tag{52}$$

where we can isolate the factor that does not depend on ρ,

$$p(\alpha|\nabla x) = \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-\frac{1}{2\epsilon^2}\gamma^2\left(1-\langle\overline{\nabla y},\overline{\nabla x}\rangle^2\right)} \underbrace{\int_0^\infty \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-\frac{1}{2\epsilon^2}\left(\rho-\gamma\langle\overline{\nabla y},\overline{\nabla x}\rangle\right)^2} \rho d\rho}_{M}. \tag{53}$$

The bracketed term M is the integral on the interval $[0, \infty)$ of a Gaussian density with mean $m \doteq \gamma\langle\overline{\nabla y}, \overline{\nabla x}\rangle = \cos(\angle\nabla y - \angle\nabla x)\|\nabla x\|$ and variance ϵ^2; it can be rewritten, using the change of variable $\xi \doteq (\rho - m)/\epsilon$, as $\int_{-m/\epsilon}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}\xi^2}\,(\epsilon\xi + m)\, d\xi$ which can be integrated by parts to yield

$$M = \frac{\epsilon e^{-\frac{1}{2}\frac{m^2}{\epsilon^2}}}{\sqrt{2\pi}} + m\left(1 - \Psi\left(-\frac{m}{\epsilon}\right)\right)$$

and therefore

$$p(\alpha|\nabla x) = \frac{1}{\sqrt{2\pi\epsilon^2}} \exp\left(-\frac{\|\nabla x\|^2 - m^2}{2\epsilon^2}\right) M \tag{54}$$

which, once written explicitly in terms of x and y, yields (14)-(15). □

Claim 5

Proof. Since $p_{\theta,G}(y, x^t)$ is sufficient for θ, and it factorizes into $\phi_{\theta,G}(y)\phi_\theta(x^t)$, then $\phi_\theta(x^t)$ is sufficient of x^t for θ. By the factorization theorem (Theorem 3.1 of Pawitan (2001)), there exist functions f_θ and ψ such that $\phi_\theta(x^t) \propto f_\theta(\psi(x^t))$, *i.e.,* the likelihood depends on the data only through the function $\psi(\cdot)$. This latter function is what is more commonly known as the sufficient statistic, which in particular has the property that $p(\theta|x^t) = p(\theta|\psi(x^t))$. However, if ϕ_θ is sufficient for θ, it is also sufficient for future data generated from θ. Formally,

$$p_G(y|x^t) = \int p_G(y|x^t, \theta)p(\theta|x^t)dP(\theta) = \int p_G(y|\theta)p(\theta|\psi(x^t))dP(\theta) = p_G(y|\psi(x^t)) \tag{55}$$

which shows that ψ minimizes the uncertainty of y for any $g \in G$ since the right-hand-side is G-invariant. The right-hand side above is the *predictive likelihood* Hinkley (1979), which must therefore be proportional to $\tilde{f}_y(\psi(x^t))$ for some $\tilde{f}$ *and the same* ψ, also by the factorization theorem. □

Claim 4

Proof. (Sketch) The integral in (38) can be split into 3 components, one of which omitted, leaving the positive component integrated on D_+, the negative component on D_-. If the distance between these two is greater than σ, however, the components are disjoint, so for each (u, v) and α, only the the positive or the negative component are non-zero, and since $\mathcal{N}$ and $\|\nabla x\|$ are both positive, and the sign is constant, rectification inside or outside the integral is equivalent. When $\sigma > d$ there is an error in the approximation, that can be bounded as a function of σ, the minimum distance and the maximum gradient component. □

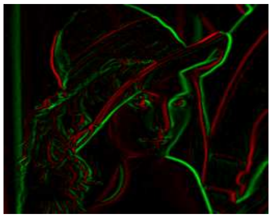

Figure 3: $D_+(0)$ *(green) and* $D_-(0)$*, the positive and negative responses to a gradient filter in the horizontal direction. The black region is their complement, which separates them.*

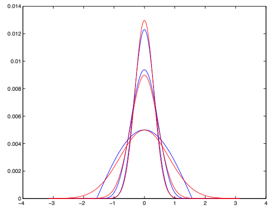

Figure 4: *Rectified cosine (blue) and its powers, compared to a Gaussian kernel (red). While the two are distinctly different for* $\epsilon = 1$*, as the power/dispersion decreases, the latter approximates the former. The plot shows* $\epsilon = 1, 1/5, 1/9$ *for the cosine, and* $1/5, 1/9, 1/13$ *for the Gaussian.*

A more general kernel could be considered, with a parameter ϵ that controls the decay, or width, of the kernel, $\kappa_\epsilon(\alpha)$. For instance, $\kappa_\epsilon(\alpha) = \kappa(\alpha)^{\frac{1}{\epsilon}}$, with the default value being $\epsilon = 1$. An alternative is to define κ to be an angular Gaussian with dispersion parameter ϵ, which is constrained to be positive and therefore does not need rectification. Although the angular Gaussian is quite different from the cosine kernel for $\epsilon = 1$, it approximates it as ϵ decreases (Fig. 4). A corollary of the above is that *the visible layer of a CNN computes the SAL Likelihood of the first hidden layer.*

The interpretation of SIFT as a likelihood function given the test image y can be confusing, as ordinarily it is interpreted as a "feature vector" associated to the training image x, and compared with other feature vectors using the Euclidean distance. In a likelihood interpretation, x is used to compute the likelihood function, and y is used to evaluate it. So, there is no descriptor built for y. The same interpretational difference applies to convolutional architectures. If interpreted as a likelihood, which would require generative learning, one would compute the likelihood of different hypotheses given the test data. Instead, currently the test data is fed to the network just as training data were, thus generating features maps, that are then compared (discriminatively) by a classifier.