A Probabilistic Framework for Deep Learning

Ankit B. Patel
Baylor College of Medicine, Rice University
ankitp@bcm.edu,abp4@rice.edu

Tan Nguyen
Rice University
mn15@rice.edu

Richard G. Baraniuk
Rice University
richb@rice.edu

Abstract

We develop a probabilistic framework for deep learning based on the *Deep Rendering Mixture Model* (DRMM), a new *generative probabilistic model* that explicitly capture variations in data due to latent task nuisance variables. We demonstrate that max-sum inference in the DRMM yields an algorithm that exactly reproduces the operations in deep convolutional neural networks (DCNs), providing a first principles derivation. Our framework provides new insights into the successes and shortcomings of DCNs as well as a principled route to their improvement. DRMM training via the Expectation-Maximization (EM) algorithm is a powerful alternative to DCN back-propagation, and initial training results are promising. Classification based on the DRMM and other variants outperforms DCNs in supervised digit classification, training 2-3$\times$ faster while achieving similar accuracy. Moreover, the DRMM is applicable to semi-supervised and unsupervised learning tasks, achieving results that are state-of-the-art in several categories on the MNIST benchmark and comparable to state of the art on the CIFAR10 benchmark.

1 Introduction

Humans are adept at a wide array of complicated sensory inference tasks, from recognizing objects in an image to understanding phonemes in a speech signal, despite significant variations such as the position, orientation, and scale of objects and the pronunciation, pitch, and volume of speech. Indeed, the main challenge in many sensory perception tasks in vision, speech, and natural language processing is a high amount of such *nuisance variation*. Nuisance variations complicate perception by turning otherwise simple statistical inference problems with a small number of variables (e.g., class label) into much higher-dimensional problems. The key challenge in developing an inference algorithm is then *how to factor out all of the nuisance variation in the input*. Over the past few decades, a vast literature that approaches this problem from myriad different perspectives has developed, but the most difficult inference problems have remained out of reach.

Recently, a new breed of machine learning algorithms have emerged for high-nuisance inference tasks, achieving super-human performance in many cases. A prime example of such an architecture is the *deep convolutional neural network* (DCN), which has seen great success in tasks like visual object recognition and localization, speech recognition and part-of-speech recognition.

The success of deep learning systems is impressive, but a fundamental question remains: *Why do they work?* Intuitions abound to explain their success. Some explanations focus on properties of feature invariance and selectivity developed over multiple layers, while others credit raw computational power and the amount of available training data. However, beyond these intuitions, a coherent theoretical framework for understanding, analyzing, and synthesizing deep learning architectures has remained elusive.

In this paper, we develop a new theoretical framework that provides insights into both the successes and shortcomings of deep learning systems, as well as a principled route to their design and improvement. Our framework is based on a *generative probabilistic model that explicitly captures variation due to latent nuisance variables*. The *Rendering Mixture Model* (RMM) explicitly models nuisance variation through a *rendering function* that combines task target variables (e.g., object class in an

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

object recognition) with a collection of task nuisance variables (e.g., pose). The *Deep Rendering Mixture Model* (DRMM) extends the RMM in a hierarchical fashion by rendering via a product of affine nuisance transformations across multiple levels of abstraction. The graphical structures of the RMM and DRMM enable efficient inference via message passing (e.g., using the max-sum/product algorithm) and training via the expectation-maximization (EM) algorithm. A key element of our framework is the relaxation of the RMM/DRMM generative model to a discriminative one in order to optimize the bias-variance tradeoff. Below, we demonstrate that the computations involved in joint MAP inference in the relaxed DRMM coincide exactly with those in a DCN.

The intimate connection between the DRMM and DCNs provides a range of new insights into how and why they work and do not work. While our theory and methods apply to a wide range of different inference tasks (including, for example, classification, estimation, regression, etc.) that feature a number of task-irrelevant nuisance variables (including, for example, object and speech recognition), for concreteness of exposition, we will focus below on the classification problem underlying visual object recognition. The proofs of several results appear in the Appendix.

2 Related Work

Theories of Deep Learning. Our theoretical work shares similar goals with several others such as the *i*-Theory [1] (one of the early inspirations for this work), Nuisance Management [27], the Scattering Transform [6], and the simple sparse network proposed by Arora et al. [2].

Hierarchical Generative Models. The DRMM is closely related to several hierarchical models, including the Deep Mixture of Factor Analyzers [30] and the Deep Gaussian Mixture Model [32].

Like the above models, the DRMM attempts to employ parameter sharing, capture the notion of nuisance transformations explicitly, learn selectivity/invariance, and promote sparsity. However, the key features that distinguish the DRMM approach from others are: (i) The DRMM explicitly models nuisance variation across multiple levels of abstraction via a product of affine transformations. This factorized linear structure serves dual purposes: it enables (ii) tractable inference (via the max-sum/product algorithm), and (iii) it serves as a regularizer to prevent overfitting by an exponential reduction in the number of parameters. Critically, (iv) inference is not performed for a single variable of interest but instead for the full global configuration of nuisance variables. This is justified in low-noise settings. And most importantly, (v) we can derive the structure of DCNs *precisely*, endowing DCN operations such as the convolution, rectified linear unit, and spatial max-pooling with principled probabilistic interpretations. Independently from our work, Soatto et al. [27] also focus strongly on nuisance management as the key challenge in defining good scene representations. However, their work considers max-pooling and ReLU as *approximations* to a marginalized likelihood, whereas our work interprets those operations differently, in terms of max-sum inference in a specific probabilistic generative model. The work on the number of linear regions in DCNs [15] is complementary to our own, in that it sheds light on the complexity of functions that a DCN can compute. Both approaches could be combined to answer questions such as: How many templates are required for accurate discrimination? How many samples are needed for learning? We plan to pursue these questions in future work.

Semi-Supervised Neural Networks. Recent work in neural networks designed for semi-supervised learning (few labeled data, lots of unlabeled data) has seen the resurgence of generative-like approaches, such as Ladder Networks [20], Stacked What-Where Autoencoders (SWWAE) [34] and many others. These network architectures augment the usual task loss with one or more regularization term, typically including an image reconstruction error, and train jointly. A key difference with our DRMM-based approach is that these networks do not arise from a proper probabilistic density and as such they must resort to learning the bottom-up recognition and top-down reconstruction weights separately, and they cannot keep track of uncertainty.

3 The Deep Rendering Mixture Model: Capturing Nuisance Variation

Although we focus on the DRMM in this paper, we define and explore several other interesting variants,including the Deep Rendering Factor Model (DRFM) and the Evolutionary DRMM (E-DRMM), both of which are discussed in more detail in [18] and the Appendix. The E-DRMM is particularly important, since its max-sum inference algorithm yields a decision tree of the type employed in a random decision forest classifier[5].

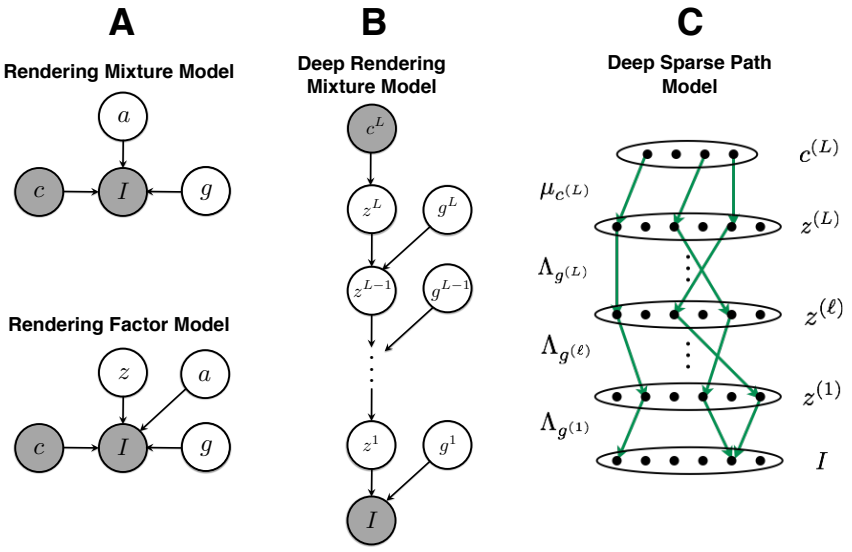

Figure 1: Graphical model depiction of (A) the Shallow Rendering Models and (B) the DRMM. All dependence on pixel location x has been suppressed for clarity. (C) The Sparse Sum-over-Paths formulation of the DRMM. A rendering path contributes only if it is active (green arrows).

3.1 The (Shallow) Rendering Mixture Model

The RMM is a *generative probabilistic model* for images that explicitly models the relationship between images I of the same object c subject to nuisance $g \in \mathcal{G}$, where $\mathcal{G}$ is the set of all nuisances (see Fig. 1A for the graphical model depiction).

$$c \sim \mathrm{Cat}(\{\pi_c\}_{c\in\mathcal{C}}), \quad g \sim \mathrm{Cat}(\{\pi_g\}_{g\in\mathcal{G}}), \quad a \sim \mathrm{Bern}(\{\pi_a\}_{a\in\mathcal{A}}),$$
$$I = a\mu_{cg} + \text{ noise}. \tag{1}$$

Here, μ_{cg} is a template that is a function of the class c and the nuisance g. The switching variable $a \in \mathcal{A} = \{\text{ON}, \text{OFF}\}$ determines whether or not to render the template at a particular patch; a sparsity prior on a thus encourages each patch to have a few causes. The noise distribution is from the exponential family, but without loss of generality we illustrate below using Gaussian noise $\mathcal{N}(0, \sigma^2\mathbf{1})$. We assume that the noise is i.i.d. as a function of pixel location x and that the class and nuisance variables are independently distributed according to categorical distributions. (Independence is merely a convenience for the development; in practice, g can depend on c.) Finally, since the world is spatially varying and an image can contain a number of different objects, it is natural to break the image up into a number of *patches*, that are centered on a single pixel x. The RMM described in (1) then applies at the patch level, where c, g, and a depend on pixel/patch location x. We will omit the dependence on x when it is clear from context.

Inference in the Shallow RMM Yields One Layer of a DCN. We now connect the RMM with the computations in one layer of a deep convolutional network (DCN). To perform object recognition with the RMM, we must marginalize out the nuisance variables g and a. Maximizing the log-posterior over $g \in \mathcal{G}$ and $a \in \mathcal{A}$ and then choosing the most likely class yields the *max-sum classifier*

$$\hat{c}(I) = \operatorname*{argmax}_{c\in\mathcal{C}} \max_{g\in\mathcal{G}} \max_{a\in\mathcal{A}} \ln p(I|c,g,a) + \ln p(c,g,a) \tag{2}$$

that computes the most likely global configuration of target and nuisance variables for the image. Assuming that Gaussian noise is added to the template, the image is normalized so that $\|I\|^2 = 1$, and c, g are uniformly distributed, (2) becomes

$$\hat{c}(I) \equiv \operatorname*{argmax}_{c\in\mathcal{C}} \max_{g\in\mathcal{G}} \max_{a\in\mathcal{A}} a(\langle w_{cg}|I\rangle + b_{cg}) + b_a \tag{3}$$
$$= \operatorname*{argmax}_{c\in\mathcal{C}} \max_{g\in\mathcal{G}} \mathrm{ReLu}(\langle w_{cg}|I\rangle + b_{cg}) + b_0 \tag{4}$$

where $\mathrm{ReLU}(u) \equiv (u)_+ = \max\{u, 0\}$ is the soft-thresholding operation performed by the rectified linear units in modern DCNs. Here we have reparameterized the RMM model from the

moment parameters $\theta \equiv \{\sigma^2, \mu_{cg}, \pi_a\}$ to the *natural parameters* $\eta(\theta) \equiv \{w_{cg} \equiv \frac{1}{\sigma^2}\mu_{cg},\ b_{cg} \equiv -\frac{1}{2\sigma^2}\|\mu_{cg}\|_2^2,\ b_a \equiv \ln p(a) = \ln \pi_a,\ b_0 \equiv \ln\left(\frac{p(a=1)}{p(a=0)}\right)$. The relationships $\eta(\theta)$ are referred to as the *generative parameter constraints*.

We now demonstrate that *the sequence of operations in the max-sum classifier in (3) coincides exactly with the operations involved in one layer of a DCN*: image normalization, linear template matching, thresholding, and max pooling. First, the image is *normalized* (by assumption). Second, the image is filtered with a set of noise-scaled rendered templates w_{cg}. If we assume *translational invariance* in the RMM, then the rendered templates w_{cg} yield a *convolutional* layer in a DCN [11] (see Appendix Lemma A.2). Third, the resulting activations (log-probabilities of the hypotheses) are passed through a pooling layer; if g is a translational nuisance, then taking the maximum over g corresponds to *max pooling* in a DCN. Fourth, since the switching variables are latent (unobserved), we max-marginalize over them during classification. This leads to the ReLU operation (see Appendix Proposition A.3).

3.2 The Deep Rendering Mixture Model: Capturing Levels of Abstraction

Marginalizing over the nuisance $g \in \mathcal{G}$ in the RMM is intractable for modern datasets, since $\mathcal{G}$ will contain all configurations of the high-dimensional nuisance variables g. In response, we extend the RMM into a hierarchical *Deep Rendering Mixture Model* (DRMM) by factorizing g into a number of different nuisance variables $g^{(1)}, g^{(2)}, \ldots, g^{(L)}$ at different levels of abstraction. The DRMM image generation process starts at the highest level of abstraction ($\ell = L$), with the random choice of the object class $c^{(L)}$ and overall nuisance $g^{(L)}$. It is then followed by random choices of the lower-level details $g^{(\ell)}$ (we absorb the switching variable a into g for brevity), progressively rendering more concrete information level-by-level ($\ell \to \ell - 1$), until the process finally culminates in a fully rendered D-dimensional image I ($\ell = 0$). Generation in the DRMM takes the form:

$$c^{(L)} \sim \text{Cat}(\{\pi_{c^{(L)}}\}),\ g^{(\ell)} \sim \text{Cat}(\{\pi_{g^{(\ell)}}\})\ \forall \ell \in [L] \tag{5}$$

$$\mu_{c^{(L)}g} \equiv \Lambda_g \mu_{c^{(L)}} \equiv \Lambda^{(1)}_{g^{(1)}} \Lambda^{(2)}_{g^{(2)}} \cdots \Lambda^{(L-1)}_{g^{(L-1)}} \Lambda^{(L)}_{g^{(L)}} \mu_{c^{(L)}} \tag{6}$$

$$I \sim \mathcal{N}(\mu_{c^{(L)}g}, \Psi \equiv \sigma^2 \mathbf{1}_D), \tag{7}$$

where the latent variables, parameters, and helper variables are defined in full detail in Appendix B.

The DRMM is a deep Gaussian Mixture Model (GMM) with special constraints on the latent variables. Here, $c^{(L)} \in \mathcal{C}^L$ and $g^{(\ell)} \in \mathcal{G}^\ell$, where $\mathcal{C}^L$ is the set of target-relevant nuisance variables, and $\mathcal{G}^\ell$ is the set of all target-irrelevant nuisance variables at level ℓ. The *rendering path* is defined as the sequence $(c^{(L)}, g^{(L)}, \ldots, g^{(\ell)}, \ldots, g^{(1)})$ from the root (overall class) down to the individual pixels at $\ell = 0$. $\mu_{c^{(L)}g}$ is the template used to render the image, and $\Lambda_g \equiv \prod_\ell \Lambda_{g^{(\ell)}}$ represents the sequence of local nuisance transformations that partially render finer-scale details as we move from abstract to concrete. Note that each $\Lambda^{(\ell)}_{g^{(\ell)}}$ is an *affine* transformation with a bias term $\alpha^{(\ell)}_{g^{(\ell)}}$ that we have suppressed for clarity. Fig. 1B illustrates the corresponding graphical model. As before, we have suppressed the dependence of $g^{(\ell)}$ on the pixel location $x^{(\ell)}$ at level ℓ of the hierarchy.

Sum-Over-Paths Formulation of the DRMM. We can rewrite the DRMM generation process by expanding out the matrix multiplications into scalar products. This yields an interesting new perspective on the DRMM, as each pixel intensity $I_x = \sum_p \lambda^{(L)}_p a^{(L)}_p \cdots \lambda^{(1)}_p a^{(1)}_p$ is the sum over all *active paths* to that pixel, of the product of weights along that path. A rendering path p is active iff every switch on the path is active i.e. $\prod_\ell a^{(\ell)}_p = 1$. While exponentially many possible rendering paths exist, only a very small fraction, controlled by the sparsity of a, are active. Fig. 1C depicts the sum-over-paths formulation graphically.

Recursive and Nonnegative Forms. We can rewrite the DRMM into a recursive form as $z^{(\ell)} = \Lambda^{(\ell+1)}_{g^{(\ell+1)}} z^{(\ell+1)}$, where $z^{(L)} \equiv \mu_{c^{(L)}}$ and $z^{(0)} \equiv I$. We refer to the helper latent variables $z^{(\ell)}$ as *intermediate rendered templates*. We also define the *Nonnegative DRMM* (NN-DRMM) as a DRMM with an extra nonnegativity constraint on the intermediate rendered templates, $z^{(\ell)} \geq 0 \forall \ell \in [L]$. The latter is enforced in training via the use of a ReLu operation in the top-down reconstruction phase of inference. Throughout the rest of the paper, we will focus on the NN-DRMM, leaving the unconstrained DRMM for future work. For brevity, we will drop the NN prefix.

Factor Model. We also define and explore a variant of the DRMM that where the top-level latent variable is Gaussian: $z^{(L+1)} \sim \mathcal{N}(0, \mathbf{1}_d) \in \mathbb{R}^d$ and the recursive generation process is otherwise

identical to the DRMM: $z^{(\ell)} = \Lambda^{(\ell+1)}_{g^{(\ell+1)}} z^{(\ell+1)}$ where $g^{(L+1)} \equiv c^{(L)}$. We call this the *Deep Rendering Factor Model* (DRFM). The DRFM is closely related to the Spike-and-Slab Sparse Coding model [25]. Below we explore some training results, but we leave most of the exploration for future work. (see Fig. 3 in Appendix C for architecture of the RFM, the shallow version of the DRFM)

Number of Free Parameters. Compared to the shallow RMM, which has $D\,|\mathcal{C}^L| \prod_\ell |\mathcal{G}^\ell|$ parameters, the DRMM has only $\sum_\ell |\mathcal{G}^{\ell+1}| D^\ell D^{\ell+1}$ parameters, an *exponential reduction in the number of free parameters* (Here $\mathcal{G}^{L+1} \equiv \mathcal{C}^L$ and D^ℓ is the number of units in the ℓ-th layer with $D^0 \equiv D$). This enables efficient inference, learning, and better generalization. Note that we have assumed dense (fully connected) Λ_g's here; if we impose more structure (e.g. translation invariance), the number of parameters will be further reduced.

Bottom-Up Inference. As in the shallow RMM, given an input image I the DRMM classifier infers the most likely global configuration $\{c^{(L)}, g^{(\ell)}\}$, $\ell = 0, 1, \ldots, L$ by executing the max-sum/product message passing algorithm in two stages: (i) bottom-up (from fine-to-coarse) to infer the overall class label $\hat{c}^{(L)}$ and (ii) top-down (from coarse-to-fine) to infer the latent variables $\hat{g}^{(\ell)}$ at all intermediate levels ℓ. First, we will focus on the fine-to-coarse pass since it leads directly to DCNs.

Using (3), the *fine-to-coarse* NN-DRMM inference algorithm for inferring the most likely cateogry $\hat{c}^L$ is given by

$$\begin{aligned}\operatorname*{argmax}_{c^{(L)}\in\mathcal{C}} \max_{g\in\mathcal{G}} \mu^T_{c^{(L)}g} I &= \operatorname*{argmax}_{c^{(L)}\in\mathcal{C}} \max_{g\in\mathcal{G}} \mu^T_{c^{(L)}} \prod_{\ell=L}^{1} \Lambda^T_{g^{(\ell)}} I \\ &= \operatorname*{argmax}_{c^{(L)}\in\mathcal{C}} \mu^T_{c^{(L)}} \max_{g^{(L)}\in\mathcal{G}^L} \Lambda^T_{g^{(L)}} \cdots \underbrace{\max_{g^{(1)}\in\mathcal{G}^1} \Lambda^T_{g^{(1)}} | I}_{\equiv I^1} \;=\; \cdots \;\equiv \operatorname*{argmax}_{c^{(L)}\in\mathcal{C}} \mu^T_{c^{(L)}} I^{(L)}. \quad (8)\end{aligned}$$

Here, we have assumed the bias terms $\alpha_{g^{(\ell)}} = 0$. In the second line, we used the max-product algorithm (distributivity of max over products i.e. for $a > 0$, $\max\{ab, ac\} = a \max\{b, c\}$). See Appendix B for full details. This enables us to rewrite (8) recursively:

$$I^{(\ell+1)} \equiv \max_{g^{(\ell+1)}\in\mathcal{G}^{\ell+1}} \underbrace{(\Lambda_{g^{(\ell+1)}})^T}_{\equiv W^{(\ell+1)}} I^{(\ell)} = \mathrm{MaxPool}(\mathrm{ReLu}(\mathrm{Conv}(I^{(\ell)}))), \quad (9)$$

where $I^{(\ell)}$ is the output *feature maps* of layer ℓ, $I^{(0)} \equiv I$ and $W^{(\ell)}$ are the filters/weights for layer ℓ. Comparing to (3), we see that the ℓ-th iteration of (8) and (9) corresponds to feedforward propagation in the ℓ-th layer of a DCN. *Thus a DCN's operation has a probabilistic interpretation as fine-to-coarse inference of the most probable configuration in the DRMM.*

Top-Down Inference. A unique contribution of our generative model-based approach is that we have a principled derivation of a top-down inference algorithm for the NN-DRMM (Appendix B). The resulting algorithm amounts to a simple top-down reconstruction term $\hat{I}_n = \Lambda_{\hat{g}_n} \mu_{\hat{c}^{(L)}_n}$.

Discriminative Relaxations: From Generative to Discriminative Classifiers. We have constructed a correspondence between the DRMM and DCNs, but the mapping is not yet complete. In particular, recall the generative constraints on the weights and biases. DCNs do not have such constraints — their weights and biases are free parameters. As a result, when faced with training data that violates the DRMM's underlying assumptions, the DCN will have more freedom to compensate. In order to complete our mapping from the DRMM to DCNs, we *relax* these parameter constraints, allowing the weights and biases to be free and independent parameters. We refer to this process as a *discriminative relaxation of a generative classifier* ([16, 4], see the Appendix D for details).

3.3 Learning the Deep Rendering Model via the Expectation-Maximization (EM) Algorithm

We describe how to learn the DRMM parameters from training data via the hard EM algorithm in Algorithm 1.

The DRMM E-Step consists of bottom-up and top-down (reconstruction) E-steps at each layer ℓ in the model. The $\gamma_{ncg} \equiv p(c, g|I_n; \theta)$ are the responsibilities, where for brevity we have absorbed a into g. The DRMM M-step consists of M-steps for each layer ℓ in the model. The per-layer M-step in turn consists of a responsibility-weighted regression, where $\mathrm{GLS}(y_n \sim x_n)$ denotes the solution to a generalized Least Squares regression problem that predict targets y_n from predictors x_n and is

Algorithm 1 Hard EM and EG Algorithms for the DRMM

E-step: $\hat{c}_n, \hat{g}_n = \underset{c,g}{\operatorname{argmax}}\ \gamma_{ncg}$

M-step: $\hat{\Lambda}_{g^{(\ell)}} = \underbrace{\text{GLS}}\left(I_n^{(\ell-1)} \sim \hat{z}_n^{(\ell)} \mid g^{(\ell)} = \hat{g}_n^{(\ell)}\right)\ \forall g^{(\ell)}$

G-step: $\Delta\hat{\Lambda}_{g^{(\ell)}} \propto \nabla_{\Lambda_{g^{(\ell)}}} \ell_{DRMM}(\theta)$

closely related to the SVD. The Iversen bracket is defined as $[\![b]\!] \equiv 1$ if expression b is true and is 0 otherwise.

There are several interesting and useful features of the EM algorithm. First, we note that it is a *derivative-free alternative to the back propagation algorithm* for training that is both intuitive and potentially much faster (provided a good implementation for the GLS problem). Second, it is easily parallelized over layers, since the M-step updates each layer separately (model parallelism). Moreover, it can be extended to a batch version so that at each iteration the model is simultaneously updated using separate subsets of the data (data parallelism). This will enable training to be distributed easily across multiple machines. In this vein, our EM algorithm shares several features with the ADMM-based Bregman iteration algorithm in [31]. However, the motivation there is from an optimization perspective and so the resulting training algorithm is not derived from a proper probabilistic density. Third, it is far more interpretable via its connections to (deep) sparse coding and to the hard EM algorithm for GMMs. The sum-over-paths formulation makes it particularly clear that the mixture components are paths (from root to pixels) in the DRMM.

G-step. For the training results in this paper, we use the Generalized EM algorithm wherein we replace the M-step with a gradient descent based G-step (see Algorithm 1). This is useful for comparison with backpropagation-based training and for ease of implementation. But before we use the G-step, we would like to make a few remarks about the proper M-step of the algorithm, saving the implementation for future work.

Flexibility and Extensibility. Since we can choose different priors/types for the nuisances g, the larger DRMM family could be useful for modeling a wider range of inputs, including scenes, speech and text. The EM algorithm can then be used to train the whole system end-to-end on different sources/modalities of labeled and unlabeled data. Moreover, the capability to sample from the model allows us to probe what is captured by the DRMM, providing us with principled ways to improve the model. And finally, in order to properly account for noise/uncertainty, it is possible in principle to extend this algorithm into a *soft* EM algorithm. We leave these interesting extensions for future work.

3.4 New Insights into Deep Convnets

DCNs are Message Passing Networks. The DRMM inference algorithm is equivalent to performing *max-sum-product message passing of the DRMM* Note that by "max-sum-product" we mean a novel combination of max-sum and max-product as described in more detail in the proofs in the Appendix. The factor graph encodes the same information as the generative model but organizes it in a manner that simplifies the definition and execution of inference algorithms [10]. Such inference algorithms are called *message passing* algorithms, because they work by passing real-valued functions called messages along the edges between nodes. In the DRMM, the messages sent from finer to coarser levels are in fact the feature maps $I^{(\ell)}$. The factor graph formulation provides a powerful interpretation: the *convolution, Max-Pooling and ReLu operations in a DCN correspond to max-sum/product inference in a DRMM*. Thus, we see that architectures and layer types commonly used in today's DCNs can be derived from precise probabilistic assumptions that entirely determine their structure. The DRMM therefore unifies two perspectives — neural network and probabilistic inference (see Table 2 in the Appendix for details).

Shortcomings of DCNs. DCNs perform poorly in categorizing transparent objects [23]. This might be explained by the fact that transparent objects generate pixels that have multiple sources, conflicting with the DRMM sparsity prior on a, which encourages few sources. DCNs also fail to classify slender and man-made objects [23]. This is because of the locality imposed by the locally-

connected/convolutional layers, or equivalently, the small size of the template $\mu_{c^{(L)}g}$ in the DRMM. As a result, DCNs fail to model long-range correlations.

Class Appearance Models and Activity Maximization. The DRMM enables us to understand how trained DCNs distill and store knowledge from past experiences in their parameters. Specifically, the DRMM generates rendered templates $\mu_{c^{(L)}g}$ via a mixture of products of affine transformations, thus implying that *class appearance models in DCNs are stored in a similar factorized-mixture form over multiple levels of abstraction.* As a result, it is the product of all the filters/weights over all layers that yield meaningful images of objects (Eq. 7). We can also shed new light on another approach to understanding DCN memories that proceeds by searching for input images that maximize the activity of a particular class unit (say, class of cats) [26], a technique we call *activity maximization*. Results from activity maximization on a high performance DCN trained on 15 million images is shown in Fig. 1 of [26]. The resulting images reveal much about how DCNs store memories. Using the DRMM, the solution $I^*_{c^{(L)}}$ of the activity maximization for class $c^{(L)}$ can be derived as the sum of individual activity-maximizing patches $I^*_{\mathcal{P}_i}$, each of which is a function of the learned DRMM parameters (see Appendix E):

$$I^*_{c^{(L)}} \equiv \sum_{\mathcal{P}_i \in \mathcal{P}} I^*_{\mathcal{P}_i}(c^{(L)}, g^*_{\mathcal{P}_i}) \propto \sum_{\mathcal{P}_i \in \mathcal{P}} \mu(c^{(L)}, g^*_{\mathcal{P}_i}). \tag{10}$$

This implies that $I^*_{c^{(L)}}$ contains multiple appearances of the same object but in various poses. Each activity-maximizing patch has its own pose $g^*_{\mathcal{P}_i}$, consistent with Fig. 1 of [26] and our own extensive experiments with AlexNet, VGGNet, and GoogLeNet (data not shown). Such images provide strong confirmational evidence that the underlying model is a mixture over nuisance parameters, as predcted by the DRMM.

Unsupervised Learning of Latent Task Nuisances. A key goal of representation learning is to disentangle the factors of variation that contribute to an image's appearance. Given our formulation of the DRMM, it is clear that DCNs are discriminative classifiers that capture these factors of variation with latent nuisance variables g. As such, the theory presented here makes a clear prediction that *for a DCN, supervised learning of task targets will lead to unsupervised learning of latent task nuisance variables*. From the perspective of manifold learning, this means that the architecture of DCNs is designed to learn and disentangle the intrinsic dimensions of the data manifolds.

In order to test this prediction, we trained a DCN to classify synthetically rendered images of naturalistic objects, such as cars and cats, with variation in factors such as location, pose, and lighting. After training, we probed the layers of the trained DCN to quantify how much linearly decodable information exists about the task target $c^{(L)}$ and latent nuisance variables g. Fig. 2 (Left) shows that the trained DCN possesses significant information about latent factors of variation and, furthermore, the more nuisance variables, the more layers are required to disentangle the factors. This is strong evidence that depth is necessary and that the amount of depth required increases with the complexity of the class models and the nuisance variations.

4 Experimental Results

We evaluate the DRMM and DRFM's performance on the MNIST dataset, a standard digit classification benchmark with a training set of 60,000 28×28 labeled images and a test set of 10,000 labeled images. We also evaluate the DRMM's performance on CIFAR10, a dataset of natural objects which include a training set of 50,000 32×32 labeled images and a test set of 10,000 labeled images. In all experiments, we use a full E-step that has a bottom-up phase and a principled top-down reconstruction phase. In order to approximate the class posterior in the DRMM, we include a Kullback-Leibler divergence term between the inferred posterior $p(c|I)$ and the true prior $p(c)$ as a regularizer [9]. We also replace the M-step in the EM algorithm of Algorithm 1 by a G-step where we update the model parameters via gradient descent. This variant of EM is known as the Generalized EM algorithm [3], and here we refer to it as EG. All DRMM experiments were done with the NN-DRMM. Configurations of our models and the corresponding DCNs are provided in the Appendix I.

Supervised Training. Supervised training results are shown in Table 3 in the Appendix. *Shallow RFM:* The 1-layer RFM (RFM sup) yields similar performance to a Convnet of the same configuration (1.21% vs. 1.30% test error). Also, as predicted by the theory of generative vs discriminative classifiers, EG training converges 2-3x faster than a DCN (18 vs. 40 epochs to reach 1.5% test error, Fig. 2, middle). *Deep RFM:* Training results from an initial implementation of the 2-layer DRFM

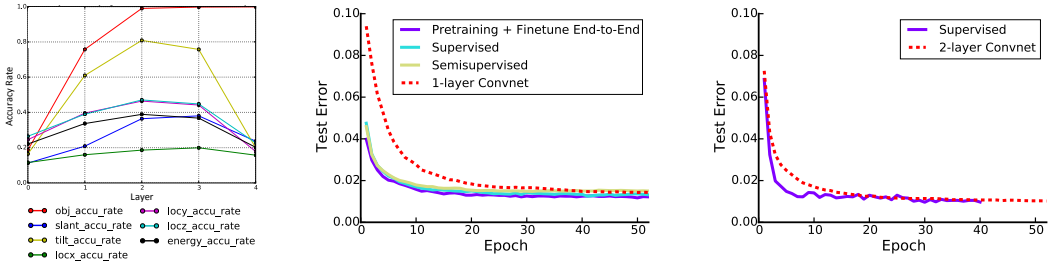

Figure 2: Information about latent nuisance variables at each layer (Left), training results from EG for RFM (Middle) and DRFM (Right) on MNIST, as compared to DCNs of the same configuration.

EG algorithm converges $2-3\times$ faster than a DCN of the same configuration, while achieving a *similar* asymptotic test error (Fig. 2, Right). Also, for completeness, we compare supervised training for a 5-layer DRMM with a corresponding DCN, and they show comparable accuracy (0.89% vs 0.81%, Table 3).

Unsupervised Training. We train the RFM and the 5-layer DRMM unsupervised with N_U images, followed by an end-to-end re-training of the whole model (unsup-pretr) using N_L labeled images. The results and comparison to the SWWAE model are shown in Table 1. The DRMM model outperforms the SWWAE model in both scenarios (Filters and reconstructed images from the RFM are available in the Appendix 4.)

Table 1: Comparison of Test Error rates (%) between best DRMM variants and other best published results on MNIST dataset for the semi-supervised setting (taken from [34]) with $N_U = 60K$ unlabeled images, of which $N_L \in \{100, 600, 1K, 3K\}$ are labeled.

Model	$N_L = 100$	$N_L = 600$	$N_L = 1K$	$N_L = 3K$
Convnet [11]	22.98	7.86	6.45	3.35
MTC [21]	12.03	5.13	3.64	2.57
PL-DAE [12]	10.49	5.03	3.46	2.69
WTA-AE [14]	-	2.37	1.92	-
SWWAE dropout [34]	8.71 ± 0.34	3.31 ± 0.40	2.83 ± 0.10	2.10 ± 0.22
M1+TSVM [8]	11.82 ± 0.25	5.72	4.24	3.49
M1+M2 [8]	3.33 ± 0.14	2.59 ± 0.05	2.40 ± 0.02	2.18 ± 0.04
Skip Deep Generative Model [13]	1.32	-	-	-
LadderNetwork [20]	1.06 ± 0.37	-	$\mathbf{0.84 \pm 0.08}$	-
Auxiliary Deep Generative Model [13]	0.96	-	-	-
catGAN [28]	1.39 ± 0.28	-	-	-
ImprovedGAN [24]	0.93 ± 0.065	-	-	-
RFM	14.47	5.61	4.67	2.96
DRMM 2-layer semi-sup	11.81	3.73	2.88	1.72
DRMM 5-layer semi-sup	3.50	**1.56**	1.67	**0.91**
DRMM 5-layer semi-sup NN+KL	**0.57**	–	–	–
SWWAE unsup-pretr [34]	-	9.80	6.135	4.41
RFM unsup-pretr	16.2	5.65	4.64	2.95
DRMM 5-layer unsup-pretr	**12.03**	**3.61**	**2.73**	**1.68**

Semi-Supervised Training. For semi-supervised training, we use a randomly chosen subset of $N_L = 100$, 600, 1K, and 3K labeled images and $N_U = 60K$ unlabeled images from the training and validation set. Results are shown in Table 1 for a RFM, a 2-layer DRMM and a 5-layer DRMM with comparisons to related work. The DRMMs performs comparably to state-of-the-art models. Specially, the 5-layer DRMM yields the best results when $N_L = 3K$ and $N_L = 600$ while results in the second best result when $N_L = 1K$. We also show the training results of a 9-layer DRMM on CIFAR10 in Table 4 in Appendix H. The DRMM yields comparable results on CIFAR10 with the best semi-supervised methods. For more results and comparisons to other related work, see Appendix H.

5 Conclusions

Understanding successful deep vision architectures is important for improving performance and solving harder tasks. In this paper, we have introduced a new family of hierarchical generative models, whose inference algorithms for two different models reproduce deep convnets and decision trees, respectively. Our initial implementation of the DRMM EG algorithm outperforms DCN back-propagation in both supervised and unsupervised classification tasks and achieves comparable/state-of-the-art performance on several semi-supervised classification tasks, with no architectural hyperparameter tuning [17, 19]

Acknowledgments. Thanks to Xaq Pitkow and Ben Poole for helpful discussions and feedback. ABP and RGB were supported by IARPA via DoI/IBC contract D16PC00003. RGB was also supported by NSF CCF-1527501, AFOSR FA9550-14-1-0088, ARO W911NF-15-1-0316, and ONR N00014-12-1-0579. TN was supported by an NSF Graduate Reseach Fellowship and NSF IGERT Training Grant (DGE-1250104).

References

[1] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio. Magic materials: a theory of deep hierarchical architectures for learning sensory representations. *MIT CBCL Technical Report*, 2013.

[2] S. Arora, A. Bhaskara, R. Ge, and T. Ma. Provable bounds for learning some deep representations. *arXiv preprint arXiv:1310.6343*, 2013.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer New York, 2006.

[4] C. M. Bishop, J. Lasserre, et al. Generative or discriminative? getting the best of both worlds. *Bayesian Statistics*, 8:3–24, 2007.

[5] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

[7] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

[8] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

[9] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[10] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, 2013.

[13] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*, 2016.

[14] A. Makhzani and B. J. Frey. Winner-take-all autoencoders. In *Advances in Neural Information Processing Systems*, pages 2773–2781, 2015.

[15] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2924–2932, 2014.

[16] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002.

[17] T. Nguyen, A. B. Patel, and R. G. Baraniuk. Semi-supervised learning with deep rendering mixture model. *CVPR(Submitted)*, 2017.

[18] A. B. Patel, T. Nguyen, and R. G. Baraniuk. A probabilistic theory of deep learning. *arXiv preprint arXiv:1504.00641*, 2015.

[19] A. B. Patel, T. Nguyen, and R. G. Baraniuk. A probabilistic framework for deep learning. *NIPS*, 2016.

[20] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3532–3540, 2015.

[21] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Muller. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, pages 2294–2302, 2011.

[22] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 833–840, 2011.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.

[25] A.-S. Sheikh, J. A. Shelton, and J. Lücke. A truncated em approach for spike-and-slab sparse coding. *Journal of Machine Learning Research*, 15(1):2653–2687, 2014.

[26] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[27] S. Soatto and A. Chiuso. Visual representations: Defining properties and deep approximations. In *International Conference on Learning Representations*, 2016.

[28] J. T. Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

[29] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[30] Y. Tang, R. Salakhutdinov, and G. Hinton. Deep mixtures of factor analysers. *arXiv preprint arXiv:1206.4635*, 2012.

[31] G. Taylor, R. Burmeister, Z. Xu, B. Singh, A. Patel, and T. Goldstein. Training neural networks without gradients: A scalable admm approach. *arXiv preprint arXiv:1605.02026*, 2016.

[32] A. van den Oord and B. Schrauwen. Factoring variations in natural images with deep gaussian mixture models. In *Advances in Neural Information Processing Systems*, pages 3518–3526, 2014.

[33] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[34] J. Zhao, M. Mathieu, R. Goroshin, and Y. LeCun. Stacked what-where autoencoders. *arXiv preprint arXiv:1506.02351*, 2016.

A From the Rendering Mixture Model Classifier to a DCN Layer

Proposition A.1 (MaxOut Neural Networks)**.** *The discriminative relaxation of a noise-free Gaussian Rendering Mixture Model (GRMM) classifier with nuisance variable $g \in \mathcal{G}$ is a single layer neural net consisting of a local template matching operation followed by a piecewise linear activation function (also known as a* MaxOut NN *[7]).*

Proof. For transparency, we prove this claim exhaustively. Later claims will have simpler proofs. We have

$$
\begin{aligned}
\hat{c}(I) &\equiv \operatorname*{argmax}_{c\in\mathcal{C}} p(c|I) \\
&= \operatorname*{argmax}_{c\in\mathcal{C}} \{p(I|c)p(c)\} \\
&= \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \sum_{g\in\mathcal{G}} p(I|c,g)p(c,g) \right\} \\
&\overset{(a)}{=} \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{g\in\mathcal{G}} p(I|c,g)p(c,g) \right\} \\
&= \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{g\in\mathcal{G}} \exp\left(\ln p(I|c,g) + \ln p(c,g)\right) \right\} \\
&\overset{(b)}{=} \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{g\in\mathcal{G}} \exp\left(\sum_{\omega} \ln p(I^\omega|c,g) + \ln p(c,g) \right) \right\} \\
&\overset{(c)}{=} \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{g\in\mathcal{G}} \exp\left(-\frac{1}{2} \sum_{\omega} \left\langle I^\omega - \mu_{cg}^\omega | \Sigma_{cg}^{-1} | I^\omega - \mu_{cg}^\omega \right\rangle + \ln p(c,g) - \frac{D}{2} \ln|\Sigma_{cg}| \right) \right\} \\
&= \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{g\in\mathcal{G}} \exp\left(\sum_{\omega} \left\langle w_{cg}^\omega | I^\omega \right\rangle + b_{cg}^\omega \right) \right\} \\
&\overset{(d)}{\equiv} \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \exp\left(\max_{g\in\mathcal{G}} \{ w_{cg} \star_{LC} I \} \right) \right\} \\
&= \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{g\in\mathcal{G}} \{ w_{cg} \star_{LC} I \} \right\} \\
&= \text{Choose}\left\{ \text{MaxOutPool}(\text{LocalTemplateMatch}(I)) \right\} \\
&= \text{MaxOut-NN}(I;\theta).
\end{aligned}
$$

In line (a), we take the noise-free limit of the GRMM, which means that one hypothesis (c, g) dominates all others in likelihood. In line (b), we assume that the image I consists of multiple channels $\omega \in \Omega$, that are conditionally independent given the global configuration (c, g) [Do we still need to consider each channel ω for the proof?]. Typically, for input images these are color channels and $\Omega \equiv \{R, G, B\}$ but in general Ω can be more abstract (e.g. as in feature maps). In line (c), we assume that the pixel noise covariance is isotropic and conditionally independent given the global configuration (c, g), so that $\Sigma_{cg} = \sigma_x^2 \mathbf{1}_D$ is proportional to the $D \times D$ identity matrix $\mathbf{1}_D$. In line (d), we defined the *locally connected template matching operator* $\star_{LC}$, which is a location-dependent template matching operation. □

Note that the nuisance variables $g \in \mathcal{G}$ are (max-)marginalized over, after the application of a local template matching operation against a set of filters/templates $\mathcal{W} \equiv \{w_{cg}\}_{c\in\mathcal{C}, g\in\mathcal{G}}$.

Lemma A.2 (**Translational Nuisance** $\rightarrow_d$ **DCN Convolution**)**.** *The MaxOut template matching and pooling operation (from Proposition A.1) for a set of translational nuisance variables $\mathcal{G} \equiv \mathcal{T}$ reduces to the traditional DCN convolution and max-pooling operation.*

Proof. Let the activation for a single output unit be $y_c(I)$. Then we have

$$\begin{aligned} y_c(I) &\equiv \max_{g\in\mathcal{G}} \{w_{cg} \star_{LC} I\} \\ &= \max_{t\in\mathcal{T}} \{\langle w_{ct}|I\rangle\} \\ &= \max_{t\in\mathcal{T}} \{\langle T_t w_c|I\rangle\} \\ &= \max_{t\in\mathcal{T}} \{\langle w_c|T_{-t}I\rangle\} \\ &= \max_{t\in\mathcal{T}} \{(w_c \star_{\text{DCN}} I)_t\} \\ &= \text{MaxPool}(w_c \star_{\text{DCN}} I). \end{aligned}$$

where $\star_{\text{DCN}}$ is the traditional DCN Convolution operator. Finally, vectorizing in c gives us the desired result $y(I) = \text{MaxPool}(\mathcal{W} \star_{\text{DCN}} I)$. □

Proposition A.3 (Max Pooling DCNs with ReLu Activations)**.** *The discriminative relaxation of a noise-free GRMM **with translational nuisances and random missing data** is a single convolutional layer of a traditional DCN. The layer consists of a generalized convolution operation, followed by a ReLu activation function and a Max-Pooling operation.*

Proof. We will model completely random missing data as a nuisance transformation $a \in \mathcal{A} \equiv \{\text{keep}, \text{drop}\}$, where $a = \text{keep} = 1$ leaves the rendered image data untouched, while $a = \text{drop} = 0$ throws out the entire image after rendering. Thus, the switching variable a models missing data. Critically, whether the data is missing is assumed to be *completely random* and thus independent of any other task variables, including the measurements (i.e. the image itself). Since the missingness of the evidence is just another nuisance, we can invoke Proposition A.1 to conclude that the discriminative relaxation of a noise-free GRMM with random missing data is also a MaxOut-DCN, but with a specialized structure which we now derive.

Mathematically, we decompose the nuisance variable $g \in \mathcal{G}$ into two parts $g = (t, a) \in \mathcal{G} = \mathcal{T} \times \mathcal{A}$, and then, following a similar line of reasoning as in Proposition A.1, we have

$$\begin{aligned} \hat{c}(I) &= \operatorname*{argmax}_{c\in\mathcal{C}} \max_{g\in\mathcal{G}} p(c, g|I) \\ &= \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{g\in\mathcal{G}} \{w_{cg} \star_{LC} I\} \right\} \\ &\overset{(a)}{=} \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{t\in\mathcal{T}} \max_{a\in\mathcal{A}} \{a(\langle w_{ct}|I\rangle + b_{ct}) + b'_{ct} + b_a + b'_I\} \right\} \\ &\overset{(b)}{=} \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{t\in\mathcal{T}} \{\max\{(w_c \star_{\text{DCN}} I)_t, 0\} + b'_{ct} + b'_{\text{drop}} + b'_I\} \right\} \\ &\overset{(c)}{=} \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{t\in\mathcal{T}} \{\max\{(w_c \star_{\text{DCN}} I)_t, 0\} + b'_{ct}\} \right\} \\ &\overset{(d)}{=} \operatorname*{argmax}_{c\in\mathcal{C}} \left\{ \max_{t\in\mathcal{T}} \{\max\{(w_c \star_{\text{DCN}} I)_t, 0\}\} \right\} \\ &= \text{Choose}\,\{\text{MaxPool}(\text{ReLu}(\text{DCNConv}(I)))\} \\ &= \text{DCN}(I; \theta). \end{aligned}$$

In line (a) we calculated the log-posterior (ignoring (c, g)-independent constants)

$$\begin{aligned} \ln p(c, g|I) &= \ln p(c, t, a|I) \\ &= \ln p(I|c, t, a) + \ln p(c, t, a) + \ln p(I) \\ &= \frac{1}{\sigma_x^2}\langle a\mu_{ct}|I\rangle - \frac{1}{2\sigma_x^2}(\|a\mu_{ct}\|_2^2 + \|I\|_2^2)) + \ln p(c, t, a) \\ &\equiv a(\langle w_{ct}|I\rangle + b_{ct}) + b'_{ct} + b_a + b'_I, \end{aligned}$$

where $a \in \{0, 1\}$, $w_{ct} \equiv \frac{1}{\sigma_x^2}\mu_{ct}$, $b_{ct} \equiv -\frac{1}{2\sigma_x^2}\|\mu_{ct}\|_2^2$, $b_a \equiv \ln p(a)$, $b'_{ct} \equiv \ln p(c, t)$, $b'_I \equiv -\frac{1}{2\sigma_x^2}\|I\|_2^2$. In line (b), we use Lemma A.2 to write the expression in terms of the DCN convolution

operator, after which we invoke the identity $\max\{u, v\} = \max\{u - v, 0\} + v \equiv \text{ReLu}(u - v) + v$ for real numbers $u, v \in \mathbb{R}$. Here we've defined $b'_{\text{drop}} \equiv \ln p(a = \text{drop})$ and we've used a slightly modified DCN convolution operator $\star_{\text{DCN}}$ defined by $w_{ct} \star_{\text{DCN}} I \equiv w_{ct} \star I + b_{ct} + \ln\left(\frac{p(a=\text{keep})}{p(a=\text{drop})}\right)$. Also, we observe that all the primed constants are independent of a and so can be pulled outside of the $\max_a$. In line(c), the two primed constants that are also independent of c, t can be dropped due to the argmax_{ct}. Finally, in line (d), we assume a uniform prior over c, t. The resulting sequence of operations corresponds *exactly* to those applied in a single convolutional layer of a traditional DCN. □

B From the Deep Rendering Mixture Model to DCNs

Here we define the DRMM in full detail.

Definition B.1 (**Deep Rendering Mixture Model (DRMM)**). *The Deep Rendering Mixture Model (DRMM) is a deep Gaussian Mixture Model (GMM) with special constraints on the latent variables. Generation in the DRMM takes the form:*

$$
\begin{aligned}
c^{(L)} &\sim \textit{Cat}(\{\pi_{c^{(L)}}\}) \\
g^{(\ell)} &\sim \textit{Cat}(\{\pi_{g^{(\ell)}}\})\ \forall \ell \in [L] \equiv \{1, 2, \ldots, L\} \\
\mu_{c^{(L)}g} &\equiv \Lambda_g \mu_{c^{(L)}} \\
&\equiv \Lambda^{(1)}_{g^{(1)}} \Lambda^{(2)}_{g^{(2)}} \ldots \Lambda^{(L-1)}_{g^{(L-1)}} \Lambda^{(L)}_{g^{(L)}} \mu_{c^{(L)}} \\
I &\sim \mathcal{N}(\mu_{c^{(L)}g}, \Psi) \\
&= \mathcal{N}(\mu_{c^{(L)}g}, \sigma^2 \boldsymbol{I}_{D^{(0)}})
\end{aligned}
$$

where the latent variables, parameters, and helper variables are defined as

$$
\begin{aligned}
g^{(\ell)} &\equiv \left(g^{(\ell)}_{x^{(\ell)}}\right)_{x^{(\ell)} \in \mathcal{X}^{(\ell)}} \\
t^{(\ell)} &\equiv \left(t^{(\ell)}_{x^{(\ell)}}\right)_{x^{(\ell)} \in \mathcal{X}^{(\ell)}},\ a^{(\ell)} \equiv \left(a^{(\ell)}_{x^{(\ell)}}\right)_{x^{(\ell)} \in \mathcal{X}^{(\ell)}} \\
g^{(\ell)}_{x^{(\ell)}} &\equiv \left(t^{(\ell)}_{x^{(\ell)}}, a^{(\ell)}_{x^{(\ell)}}\right) \\
t^{(\ell)}_{x^{(\ell)}} &\in \{\textit{UL}, \textit{UR}, \textit{LL}, \textit{LR}\} \\
a^{(\ell)}_{x^{(\ell)}} &\in \{0, 1\} \equiv \{\textit{OFF}, \textit{ON}\} \\
x^{(\ell)} &\in \mathcal{X}^{(\ell)} \equiv \{\textit{pixels in level } \ell\} \in \mathbb{R}^{D^{(\ell)}} \\
\Lambda^{(\ell)}_{g^{(\ell)}} &= \Lambda^{(\ell)}_{t^{(\ell)}, a^{(\ell)}} \in \mathbb{R}^{D^{(\ell-1)} \times D^{(\ell)}} \\
&= T^{(\ell)}_{t^{(\ell)}} Z^{(\ell)} \Gamma^{(\ell)} M^{(\ell)}_{a^{(\ell)}} \\
M^{(\ell)}_{a^{(\ell)}} &\equiv \textit{diag}\left(a^{(\ell)}\right) \in \mathbb{R}^{D^{(\ell)} \times D^{(\ell)}} \\
T^{(\ell)}_{t^{(\ell)}} &\equiv \textit{translation operator to position } t^{(\ell)} \in \mathbb{R}^{D^{(\ell-1)} \times D^{(\ell-1)}} \\
Z^{(\ell)} &\equiv \textit{zero-padding operator} \in \mathbb{R}^{D^{(\ell-1)} \times F^{(\ell)}} \\
\Gamma^{(\ell)} &\equiv \bigotimes_{x^{(\ell)} \in \mathcal{X}^{(\ell)}} \underbrace{\Gamma^{(\ell)}_{x^{(\ell)}}}_{F^{(\ell)} \times 1} \in \mathbb{R}^{F^{(\ell)} \times D^{(\ell)}} \\
\Gamma^{(\ell)}_{x^{(\ell)}} &\equiv \{\textit{filter bank at level } \ell\} \in \mathbb{R}^{F^{(\ell)}} \\
F^{(\ell)} &\equiv W^{(\ell)} H^{(\ell)} C^{(\ell)} \\
&= \textit{ size of the core templates at layer } (\ell)
\end{aligned}
$$

For simplicity, in the following sections, we will use c and $c^{(L)}$ interchangeably.

Definition B.2 (Nonnegative Deep Rendering Mixture Model (NN-DRM)). *The Nonnegative Deep Rendering Mixture Model is defined as a DRMM (Definition B.1) with additional nonnegativity constraint(s) on the intermediate latent variables (rendered templates):*

$$z_n^{(\ell)} = \Lambda_{g_n^{(\ell+1)}} \cdots \Lambda_{g_n^{(L)}} \mu_{c_n^{(L)}} \geq 0 \quad \forall \ell \in \{1, \ldots, L\} \tag{11}$$

Following the same line of reasoning as in the main text, we will derive the Hard EM algorithm for the DRMM model.

B.1 E-step: Computing the Soft Responsibilities

$$\begin{aligned}
\gamma_{ncg} &\equiv p(c, g|I_n) \\
&= \frac{p(I_n|c, g; \theta)p(c, g|\theta)}{\sum_{c,g} p(I_n|c, g; \theta)p(c, g|\theta)} \\
&= \frac{\pi_{cg}|\Psi|^{-1/2} \exp\left(-\frac{1}{2}\|I_n - \mu_{cg}\|^2_{\Psi^{-1}}\right)}{Z},
\end{aligned}$$

where the partition function Z is defined as

$$Z(\theta) \equiv \sum_{c,g} \pi_{cg}|\Psi|^{-1/2} \exp\left(-\frac{1}{2}\|I_n - \mu_{cg}\|^2_{\Psi^{-1}}\right).$$

Since the numerator and denominator both contain $|\Psi|^{-1/2}$, the responsibilities simplify to

$$\gamma_{ncg} = \frac{\pi_{cg} \exp\left(-\frac{1}{2}\|I_n - \mu_{cg}\|^2_{\Psi^{-1}}\right)}{Z'}, \tag{12}$$

where Z' is defined as

$$Z'(\theta) \equiv \sum_{cg} \pi_{cg} \exp\left(-\frac{1}{2}\|I_n - \mu_{cg}\|^2_{\Psi^{-1}}\right).$$

B.2 E-step: Computing the Hard Responsibilities

Assuming isotropic noise $\Psi = \sigma^2 1_D$ and taking the zero-noise limit $\sigma^2 \to 0$, the term in the denominator $Z'(\theta)$ for which $\|I_n - \mu_{cg}\|_2^2$ is smallest will go to zero most slowly. Hence the responsibilities γ_{ncg} will all approach zero, except for one term (c^*, g^*), for which the $\gamma_{nc^*g^*}$ will approach one. [1] Thus, the soft responsibilities become hard responsibilities in the zero-noise limit:

$$\gamma_{ncg} \xrightarrow{\sigma \to 0} r_{ncg} \equiv \begin{cases} 1, & \text{if } (c, g) = \operatorname{argmax}_{c'g'} -\frac{1}{2}\|I_n - \mu_{c'g'}\|_2^2 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

B.3 Useful Lemmas

In order to derive the E-step for the DRMM, we will need a few simple theoretical results. We prove them here.

Definition B.3 (Masking Operator). *Let $a \in \{0, 1\}^d$ be a binary vector (mask) and let $\Lambda \in \mathbb{R}^{D \times d}$ be a real matrix. Then the **masking operator** $\mathcal{M}_a(\Lambda) \in \mathbb{R}^{D \times d}$ is defined as*

$$\mathcal{M}_a(\Lambda) \equiv \Lambda \cdot M_a \equiv \Lambda \cdot \operatorname{diag}(a),$$

where $M_a \equiv \operatorname{diag}(a) \in \mathbb{R}^{d \times d}$ is the diagonal masking matrix.

[1] Technically, there can be multiple maximizers and the algorithms below can be generalized to handle this case. But we focus on the case with just one unique maximum for simplicity.

Lemma B.4. *The action of a masking operator on a vector $z \in \mathbb{R}^d$ can be written in several equivalent ways:*

$$\begin{aligned}\mathcal{M}_a(\Lambda)z &= \Lambda \cdot \mathrm{diag}(a) \cdot z \\ &= \Lambda \cdot \mathrm{diag}(a) \cdot \mathrm{diag}(a) \cdot z \\ &= \Lambda[:, a] \cdot z[a] \\ &= \Lambda(a \odot z).\end{aligned}$$

Here $\odot$ denotes the elementwise (Hadamard) product between two vectors and $\Lambda[:, a]$ is numpy notation for the subset of columns $\{j \in [D] : a_j = 1\}$ of Λ.

Proof. The first equality is by definition. The second equality is a result of a being binary since $a_i^2 = a_i$ for $a_i \in \{0, 1\}$. The third and fourth equalities result from the associativity of matrix multiplication. □

Lemma B.5 (Optimization with Masking Operators). *Let $z, u \in \mathbb{R}^{D\times 1}$. Consider the optimization problem*

$$\max_{a\in\{0,1\}^D} \mathcal{M}_a(z^T)u = \max_{a\in\{0,1\}^D} z^T M_a u \tag{14}$$

where $M_a \equiv \mathrm{diag}(a)$. Then the optimization can be solved in closed form as:

(a) $\max_{a\in\{0,1\}^D} \mathcal{M}_a(z^T)u = \mathbf{1}_D^T \,\mathrm{ReLu}(z \odot u)$.

(b) $\hat{a} \equiv \operatorname{argmax}_{a\in\{0,1\}^D} \mathcal{M}_a(z^T)u = [z \odot u > 0] \in \{0, 1\}^D$.

(c) $M_{\hat{a}}u = \mathrm{sgn}(z) \odot \mathrm{ReLu}\left(\mathrm{sgn}(z) \odot u\right)$.

(d) If $z \geq 0$, then $\hat{a} \equiv \operatorname{argmax}_{a\in\{0,1\}^D} \mathcal{M}_a(z^T)u = [u > 0] \in \{0, 1\}^D$ is a maximizer, for which $M_{\hat{a}}u = \mathrm{ReLu}\left(u\right)$.

Proof. (a) The maximum value can be computed as

$$\begin{aligned}v^\star &\equiv \max_{a\in\{0,1\}^D} \mathcal{M}_a(z^T)u \\ &= \max_{a\in\{0,1\}^D} z^T \,\mathrm{diag}(a)u \\ &= \max_{a\in\{0,1\}^D} \sum_{i\in[D]} z_i a_i u_i \\ &= \sum_{i\in[D]} \max_{a_i\in\{0,1\}} a_i(z_i u_i) \\ &\equiv \sum_{i\in[D]} \hat{a}_i(z_i u_i) \\ &= \sum_{i\in[D]} [z_i u_i > 0] \cdot z_i u_i \\ &= \sum_{i\in[D]} \mathrm{ReLu}(z_i u_i) \\ &= \mathbf{1}_D^T \,\mathrm{ReLu}(z \odot u).\end{aligned}$$

(b) In the 4th line the vector optimization decouples into a set of independent scalar optimizations $\max_{a_i\in\{0,1\}} a_i(z_i u_i)$, each of which is solvable in closed form: $\hat{a}_i \equiv [z_i u_i > 0]$. Hence, the optimal

solution $\hat{a}$ is given by $\hat{a} = [z \odot u > 0]$.
(c) Substituting in $\hat{a}$, we get

$$\begin{aligned}
\mathcal{M}_{\hat{a}} u &= u \odot [z \odot u > 0] \\
&= \underbrace{(\operatorname{sgn}(z) \odot \operatorname{sgn}(z))}_{\mathbf{1}_D} \odot u \odot [\operatorname{sgn}(z) \odot u > 0] \\
&= \operatorname{sgn}(z) \odot (\operatorname{sgn}(z) \odot u) \odot [\operatorname{sgn}(z) \odot u > 0] \\
&= \operatorname{sgn}(z) \odot \operatorname{ReLu}(\operatorname{sgn}(z) \odot u),
\end{aligned}$$

where in the third and fourth equalities we have used the associativity of elementwise multiplication and the definition of ReLu, respectively.
(d) If $z \geq 0$, then when $z_i > 0$, $\hat{a}_i = [u_i > 0]$, and when $z_i = 0$, $\hat{a}_i$ can be either 0 or 1 since then $\max_{a_i \in \{0,1\}} a_i(z_i u_i) = 0 \, \forall a_i \in \{0, 1\}$. Therefore, if $z \geq 0$, $\hat{a} = [u > 0]$ is a solution of the optimization 14. It follows that $M_{\hat{a}} u = [u > 0]u = \operatorname{ReLu}(u)$.

□

Lemma B.6 (Optimization with "Row" Max-Marginal). *Let $z, u \in \mathbb{R}^{D \times 1}$. Consider the optimization problem*

$$\max_{t \in \mathcal{T}^D} z^T u(t) = \max_{t \in \mathcal{T}^D} \sum_x z_x u(t)_x \tag{15}$$

*where $\mathcal{T}$ is the set of possible **fine-scale** translations at location x. Also,*

$$t \equiv \begin{bmatrix} \vdots \\ t_x \\ \vdots \end{bmatrix} \text{ and } u(t) \equiv \begin{bmatrix} \vdots \\ u_x(t_x) \\ \vdots \end{bmatrix} \tag{16}$$

Then the optimization can be solved as:

(a) $\max_{t \in \mathcal{T}^D} z^T u(t) = \sum_x |z_x| \max_{t_x \in \mathcal{T}} \operatorname{sgn}(z_x) u_x(t_x)$

(b) $\hat{t} = \operatorname*{argmax}_{t \in \mathcal{T}^D} z^T u(t) = \operatorname*{argmax}_t \operatorname{sgn}(z) \odot u(t) = \begin{bmatrix} \vdots \\ \operatorname*{argmax}_{t_x} \operatorname{sgn}(z_x) u_x(t_x) \\ \vdots \end{bmatrix}$

(c) $u(\hat{t}) = \operatorname{sgn}(z) \odot \max_t (\operatorname{sgn}(z) \odot u(t)) = \begin{bmatrix} \vdots \\ \operatorname{sgn}(z_x) \max_{t_x} \operatorname{sgn}(z_x) u_x(t_x) \\ \vdots \end{bmatrix}$

(d) If $z \geq 0$, then $\hat{t} = \operatorname*{argmax}_t u(t) = \begin{bmatrix} \vdots \\ \operatorname*{argmax}_{t_x} u_x(t_x) \\ \vdots \end{bmatrix}$ is a maximizer for which $u(\hat{t}) = \max_t (u(t)) = \begin{bmatrix} \vdots \\ \max_{t_x} u_x(t_x) \\ \vdots \end{bmatrix}$

Proof. *(a)* The maximum value can be computed as

$$\begin{aligned} v^\star &\equiv \max_{\{t_x \in \mathcal{T}\}_{x=1}^{D}} \sum_x z_x u_x(t_x) \\ &= \sum_x \max_{t_x \in \mathcal{T}} z_x u_x(t_x) \\ &= \sum_x \max_{t_x \in \mathcal{T}} |z_x| \operatorname{sgn}(z_x) u_x(t_x) \\ &= \sum_x |z_x| \max_{t_x \in \mathcal{T}} \operatorname{sgn}(z_x) u_x(t_x) \end{aligned}$$

(b) In the 2nd line the vector optimization decouples into a set of independent scalar optimizations $\max_{t_x \in \mathcal{T}} z_x u_x(t_x)$, each of which has the solution as follows: $\operatorname*{argmax}_{t_x} \operatorname{sgn}(z_x) u_x(t_x)$. Hence, the optimal solution $\hat{t} = \operatorname*{argmax}_{t \in \mathcal{T}^D} z^T u(t) = \begin{bmatrix} \vdots \\ \operatorname*{argmax}_{t_x} \operatorname{sgn}(z_x) u_x(t_x) \\ \vdots \end{bmatrix} = \operatorname*{argmax}_t \operatorname{sgn}(z) \odot u(t).$

(c) Substituting in $\hat{t}$, we obtain

$$\begin{aligned} v^\star &= \sum_x |z_x| \operatorname{sgn}(z_x) u_x(\hat{t}_x) \\ &= \sum_x z_x \operatorname{sgn}(z_x) (\operatorname{sgn}(z_x) u_x(\hat{t}_x)) \\ &= \sum_x z_x \operatorname{sgn}(z_x) \max_{t_x} \operatorname{sgn}(z_x) u_x(t_x) \\ &= z^T \begin{bmatrix} \vdots \\ \operatorname{sgn}(z_x) \max_{t_x} \operatorname{sgn}(z_x) u_x(t_x) \\ \vdots \end{bmatrix} \end{aligned}$$

Hence,

$$u(\hat{t}) = \begin{bmatrix} \vdots \\ \operatorname{sgn}(z_x) \max_{t_x} \operatorname{sgn}(z_x) u_x(t_x) \\ \vdots \end{bmatrix} = \operatorname{sgn}(z) \odot \max_t \left(\operatorname{sgn}(z) \odot u(t) \right),$$

(d) If $z \geq 0$, then when $z_i > 0$, $\hat{t}_x = \operatorname*{argmax}_{t_x} u_x(t_x)$, and when $z_i = 0$, $\hat{t}_x$ can take any value in its domain since then $\max_{t_x \in \mathcal{T}} \operatorname{sgn}(z_x) u_x(t_x) = 0 \, \forall t_x \in \mathcal{T}$. Therefore, if $z \geq 0$, $\hat{t}_x = \operatorname*{argmax}_{t_x} u_x(t_x)$ is a solution of the optimization 15. It follows that $u(\hat{t}) = \max_t (u(t)) \equiv \begin{bmatrix} \vdots \\ \max_{t_x} u_x(t_x) \\ \vdots \end{bmatrix}$. □

Definition B.7 (Deep Masking Operator). *Let $a^{(\ell)} \in \{0,1\}^{D^{(\ell)}}$ be a collection of binary (vector) masks and let $\Lambda^{(\ell)} \in \mathbb{R}^{D^{(\ell-1)} \times D^{(\ell)}}$ be a collection of (matrix) operators. Then the **deep masking operator** $\mathcal{M}_{\{a^{(\ell)}\}}(\{\Lambda^{(\ell)}\}) \in \mathbb{R}^{D^{(0)} \times D^{(L)}}$ is defined as*

$$\mathcal{M}_{\{a^{(\ell)}\}}(\{\Lambda^{(\ell)}\}) \equiv \prod_{\ell=1}^{L} \mathcal{M}_{a^{(\ell)}}(\Lambda^{(\ell)}) = \prod_{\ell=1}^{L} \Lambda^{(\ell)} \cdot M_{a^{(\ell)}},$$

where $M_a \equiv \operatorname{diag}(a)$ is the diagonal masking matrix for mask a.

B.4 E-Step: Inference of Top-Level Category

Theorem B.8 (Inference in DRMM ⇒ Signed Convnets). *Inference in the DRMM, according to the Dynamic Programming-based algorithm below, yields Signed DCNs. The inference algorithm has a bottom-up and top-down pass.*

Proof. Given input image $I_n \equiv I_n^{(0)}$, we infer $\hat{c}_n$ as follows:

$$\begin{aligned}
\hat{c}_n &= \operatorname*{argmax}_c \max_g -\frac{1}{2}\|I_n - \mu_{cg}\|_2^2 \\
&= \operatorname*{argmax}_c \max_g \mu_{cg}^T I_n - \frac{1}{2}\|I_n\|_2^2 - \frac{1}{2}\|\mu_{cg}\|_2^2 \\
&= \operatorname*{argmax}_c \max_g \mu_{cg}^T I_n - \frac{1}{2}\|\mu_{cg}\|_2^2,
\end{aligned}$$

where the last equality follow since I_n is independent of c, g. We further assume that:

$$\begin{aligned}
\alpha_{g^{(\ell)}} &= 0 \ \forall \ell \\
\|\mu_{cg}\|_2^2 &= \text{const} \ \ \forall c, g.
\end{aligned}$$

As a result, $\mu_{cg} = \Lambda_g \mu_c$ and the most probable class $\hat{c}_n$ is inferred as

$$\hat{c}_n = \operatorname*{argmax}_c \max_g \mu_{cg}^T I_n^{(0)} \tag{17}$$

$$= \operatorname*{argmax}_c \max_g (\Lambda_g \mu_c)^T I_n^{(0)} \tag{18}$$

$$= \operatorname*{argmax}_c \max_{g^{(L:1)}} \mu_c^T \Lambda_{g^{(L)}}^T \cdots \Lambda_{g^{(2)}}^T \Lambda_{g^{(1)}}^T I_n^{(0)} \tag{19}$$

$$= \operatorname*{argmax}_c \max_{g^{(L:2)}} \max_{t^{(1)}} \max_{a^{(1)}} \mu_c^T \Lambda_{g^{(L)}}^T \cdots \Lambda_{g^{(2)}}^T (M_{a^{(1)}} \Lambda_{t^{(1)}}^T) I_n^{(0)} \tag{20}$$

$$= \operatorname*{argmax}_c \max_{g^{(L:2)}} \max_{t^{(1)}} \max_{a^{(1)}} \underbrace{\left(\mu_c^T \Lambda_{g^{(L)}}^T \cdots \Lambda_{g^{(2)}}^T\right)}_{\equiv z^{(1)\downarrow T}} M_{a^{(1)}} \underbrace{\left(\Lambda_{t^{(1)}}^T I_n^{(0)}\right)}_{\equiv u_n^{(1)\uparrow}(t^{(1)})} \tag{21}$$

$$= \operatorname*{argmax}_c \max_{g^{(L:2)}} \max_{t^{(1)}} \max_{a^{(1)}} z^{(1)\downarrow T} M_{a^{(1)}} u_n^{(1)\uparrow}(t^{(1)}) \tag{22}$$

$$\overset{(a)}{=} \operatorname*{argmax}_c \max_{g^{(L:2)}} \max_{t^{(1)}} z^{(1)\downarrow T} M_{\hat{a}_n^{(1)}} u_n^{(1)\uparrow}(t^{(1)}) \tag{23}$$

$$\overset{(b)}{=} \operatorname*{argmax}_c \max_{g^{(L:2)}} z^{(1)\downarrow T} \left(s^{(1)\downarrow} \odot \max_{t^{(1)}} s^{(1)\downarrow} \odot \left(M_{\hat{a}_n^{(1)}} u_n^{(1)\uparrow}(t^{(1)})\right)\right) \tag{24}$$

$$\overset{(c)}{=} \operatorname*{argmax}_c \max_{g^{(L:2)}} z^{(1)\downarrow T} \left(s^{(1)\downarrow} \odot \max_{t^{(1)}} s^{(1)\downarrow} \odot \left(s^{(1)\downarrow} \odot \operatorname{ReLu}\left(s^{(1)\downarrow} \odot u_n^{(1)\uparrow}(t^{(1)})\right)\right)\right) \tag{25}$$

$$\overset{(d)}{=} \operatorname*{argmax}_c \max_{g^{(L:2)}} z^{(1)\downarrow T} \underbrace{\left(s^{(1)\downarrow} \odot \operatorname{MaxPool}\left(\operatorname{ReLu}\left(\operatorname{diag}(s^{(1)\downarrow}) u_n^{(1)\uparrow}(\mathcal{T})\right)\right)\right)}_{\equiv I_n^{(1)}(s^{(1)\downarrow})} \tag{26}$$

$$= \operatorname*{argmax}_c \max_{g^{(L:2)}} \mu_c^T \Lambda_{g^{(L)}}^T \cdots \Lambda_{g^{(2)}}^T I_n^{(1)} \tag{27}$$

In line (a), we employ Lemma B.5*(b)* to infer the optimal $\hat{a}_n^{(1)}$. In line (b) and (c), we employ B.6*(c)* and Lemma B.5*(c)* to calculate the max-product message $I_n^{(1)}$ to be sent to the next layer. Notice that here $s^{(1)\downarrow} = \operatorname{sgn}\left(z^{(1)\downarrow}\right)$. In line (b), $\hat{t}_n^{(1)}$ is implicitly inferred via Lemma B.6*(b)*. In line (d), $s^{(1)\downarrow} \odot s^{(1)\downarrow}$ becomes a vector of all 1's. Also, in the same line, $\operatorname{diag}(s^{(1)\downarrow})$ is a diagonal matrix with diagonal $s^{(1)\downarrow}$ and $u_n^{(1)\uparrow}(\mathcal{T})$ is a matrix $[u_{nxt}]$ where rows are indexed by $x \in \mathcal{X}$ and columns by $t \in \mathcal{T}$. It corresponds to the output of the convolutional layer in a DCN, prior to the ReLu and spatial max-pooling operators.

Note that we have succeeded in expressing the optimization (Eq. 19) recursively in terms of a one level smaller sub-problem (Eq. 27). Iterating this procedure yields a set of recurrence relations, which define our *Dynamic Programming (DP)* algorithm for the bottom-up and top-down inference in the DRMM:

Bottom-Up E-Step ($\mathrm{E}_\uparrow$):

$$u_n^{(\ell)\uparrow} = \Lambda_{t^{(\ell)}}^T I_n^{(\ell-1)} \tag{28}$$

$$s^{(\ell)\downarrow} = \operatorname{sgn}\left(z^{(\ell)\downarrow}\right) \tag{29}$$

$$\forall s^{(\ell)\downarrow} \in \{\pm 1\}^{D^{(\ell)}} : \hat{a}_n^{(\ell)\updownarrow}(s^{(\ell)\downarrow}) = [s^{(\ell)\downarrow} \odot u_n^{(\ell)\uparrow} > 0] \tag{30}$$

$$\forall s^{(\ell)\downarrow} \in \{\pm 1\}^{D^{(\ell)}} : \hat{t}_n^{(\ell)\updownarrow}(s^{(\ell)\downarrow}) = \operatorname*{argmax}_{t^{(\ell)}} s^{(\ell)\downarrow} \odot u_n^{(\ell)\uparrow}(t^{(\ell)}) \tag{31}$$

$$I_n^{(\ell)}(s^{(\ell)\downarrow}) = M_{\hat{a}_n^{(\ell)}}\left(\Lambda_{\hat{t}^{(\ell)}}^T I_n^{(\ell-1)}\right) \tag{32}$$

$$= s^{(\ell)\downarrow} \odot \operatorname{MaxPool}\left(\operatorname{ReLu}\left(\operatorname{diag}(s^{(1)\downarrow}) u_n^{(1)\uparrow}(\mathcal{T})\right)\right) \tag{33}$$

$$\hat{c}_n^{(L)} = \operatorname*{argmax}_{c^{(L)}} \mu_{c^{(L)}}^T I_n^{(L)} \tag{34}$$

Top-Down/Traceback E-Step ($\mathrm{E}_\uparrow$):

$$\hat{z}_n^{(\ell)\downarrow} = \Lambda_{\hat{g}_n^{(\ell+1)}} \cdots \Lambda_{\hat{g}_n^{(L)}} \mu_{\hat{c}_n^{(L)}} \tag{35}$$

$$= \Lambda_{\hat{g}_n^{(\ell+1)}} \hat{z}_n^{(\ell+1)\downarrow} \tag{36}$$

$$\hat{s}_n^{(\ell)\downarrow} = \operatorname{sgn}(\hat{z}_n^{(\ell)\downarrow}) \tag{37}$$

$$\hat{a}_n^{(\ell)\updownarrow} = \hat{a}_n^{(\ell)\updownarrow}(\hat{s}_n^{(\ell)\downarrow}) = [\hat{s}_n^{(\ell)\downarrow} \odot u_n^{(\ell)\uparrow} > 0] \tag{38}$$

$$\hat{t}_n^{(\ell)\updownarrow} = \hat{t}_n^{(\ell)\updownarrow}(s_n^{(\ell)\downarrow}) = \operatorname*{argmax}_{t^{(\ell)}} s_n^{(\ell)\downarrow} \odot u_n^{(\ell)\uparrow}(t^{(\ell)}) \tag{39}$$

where $u_n^{(\ell)\uparrow}$ and $\hat{z}_n^{(\ell)\downarrow}$ are the bottom-up and top-down net inputs into layer ℓ, respectively. □

Corollary B.9 (Inference in the NN-DRMM ⇒ Convnets). *Inference in the NN-DRMM according to the Dynamic Programming-based algorithm above yields ReLu DCNs.*

Proof. The NN-DRMM assumes that the intermediate rendered latent variables $z_n^{(\ell)} \geq 0$ for all ℓ, which implies that the signs are also nonnegative i.e., $s_n^{(\ell)} \geq 0$. This in turn, according to Lemma B.5*(d)* and B.6*(d)*, reduces Eqs. 33, 34, 38 and 39 to

$$\mathrm{E}_\uparrow : I_n^{(\ell)} = \operatorname{MaxPool}\operatorname{ReLu}\left(u_n^{(\ell)\uparrow}\right) \tag{40}$$

$$\hat{c}_n^{(L)} = \operatorname*{argmax}_{c^{(L)}} \mu_{c^{(L)}}^T I_n^{(L)} \tag{41}$$

$$\mathrm{E}_\downarrow : \hat{a}_n^{(\ell)} = [u_n^{(\ell)\uparrow} > 0] \tag{42}$$

$$\hat{t}_n^{(\ell)} = \operatorname*{argmax}_{t^{(\ell)}} u_n^{(\ell)\uparrow}(t^{(\ell)}), \tag{43}$$

which is equivalent to feedforward propagation in a DCN. Note that the the top-down step no longer requires information from the deeper levels, and so it can be computed in the bottom-up step instead. □

Remark: Note that the vector max notation $\max_t u(t) = \begin{bmatrix} \vdots \\ \max_{t_x} u_x(t_x) \\ \vdots \end{bmatrix}$ is the same as the max notation we use in our arXiv post. It refers to the **row** max-marginals of the matrix $u(t) \equiv [u_{xt}]_{x\in\mathcal{X}, t\in\mathcal{T}}$ with respect to latent variables t.

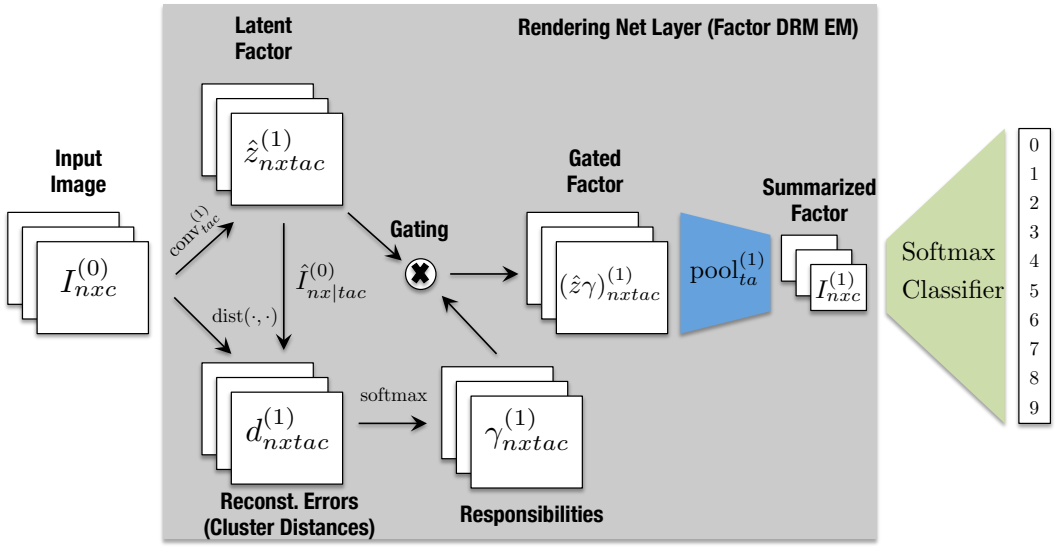

Figure 3: Neural network implementation of shallow Rendering Model EM algorithm.

C Rendering Factor Model (RFM) Architecture

D Transforming a Generative Classifier into a Discriminative One

Before we formally define the procedure, some preliminary definitions and remarks will be helpful. A generative classifier models the joint distribution $p(c, I)$ of the input features *and* the class labels. It can then classify inputs by using Bayes Rule to calculate $p(c|I) \propto p(c, I) = p(I|c)p(c)$ and picking the most likely label c. Training such a classifier is known as *generative learning*, since one can generate synthetic features I by sampling the joint distribution $p(c, I)$. Therefore, a generative classifier learns an *indirect* map from input features I to labels c by modeling the joint distribution $p(c, I)$ of the labels *and* the features.

In contrast, a discriminative classifier parametrically models $p(c|I) = p(c|I; \theta_d)$ and then trains on a dataset of input-output pairs $\{(I_n, c_n)\}_{n=1}^N$ in order to estimate the parameter θ_d. This is known as *discriminative learning*, since we directly discriminate between different labels c given an input feature I. Therefore, a discriminative classifier learns a direct map from input features I to labels c by *directly* modeling the conditional distribution $p(c|I)$ of the labels *given* the features.

Given these definitions, we can now define the *discriminative relaxation* procedure for converting a generative classifier into a discriminative one. Starting with the standard learning objective for a generative classifier, we will employ a series of transformations and relaxations to obtain the learning

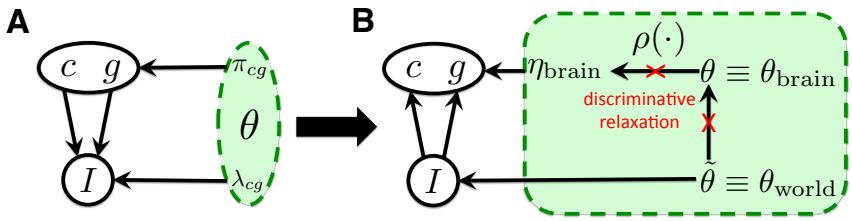

Figure 4: Graphical depiction of discriminative relaxation procedure. (A) The Rendering Model (RM) is depicted graphically, with mixing probability parameters π_{cg} and rendered template parameters λ_{cg}. Intuitively, we can interpret the discriminative relaxation as a *brain-world transformation* applied to a generative model. According to this interpretation, instead of the world generating images and class labels (A), we instead imagine the world generating images I_n via the rendering parameters $\tilde{\theta} \equiv \theta_{\text{world}}$ while the brain generates labels c_n, g_n via the classifier parameters $\eta_{\text{dis}} \equiv \eta_{\text{brain}}$ (B). The brain-world transformation converts the RM (A) to an equivalent graphical model (B), where an extra set of parameters $\tilde{\theta}$ and constraints (arrows from θ to $\tilde{\theta}$ to η) have been introduced. Discriminatively relaxing these constraints (B, red X's) yields the single-layer DCN as the discriminative counterpart to the original generative RM classifier in (A).

objective for a discriminative classifier. Mathematically, we have

$$
\begin{aligned}
\max_{\theta} L_{\text{gen}}(\theta) &\equiv \max_{\theta} \sum_{n} \ln p(c_n, I_n|\theta) \\
&\overset{(a)}{=} \max_{\theta} \sum_{n} \ln p(c_n|I_n, \theta) + \ln p(I_n|\theta) \\
&\overset{(b)}{=} \max_{\theta, \tilde{\theta}:\theta=\tilde{\theta}} \sum_{n} \ln p(c_n|I_n, \theta) + \ln p(I_n|\tilde{\theta}) \\
&\overset{(c)}{\leq} \max_{\theta} \underbrace{\sum_{n} \ln p(c_n|I_n, \theta)}_{\equiv L_{\text{cond}}(\theta)} \\
&\overset{(d)}{=} \max_{\eta:\eta=\rho(\theta)} \sum_{n} \ln p(c_n|I_n, \eta) \\
&\overset{(e)}{\leq} \max_{\eta} \underbrace{\sum_{n} \ln p(c_n|I_n, \eta)}_{\equiv L_{\text{dis}}(\eta)},
\end{aligned}
\tag{44}
$$

where the L's are the *generative*, *conditional* and *discriminative* log-likelihoods, respectively. In line (a), we used the Chain Rule of Probability. In line (b), we introduced an extra set of parameters $\tilde{\theta}$ while also introducing a constraint that enforces equality with the old set of generative parameters θ. In line (c), we relax the equality constraint (first introduced by Bishop, LaSerre and Minka in [4]), allowing the classifier parameters θ to differ from the image generation parameters $\tilde{\theta}$. In line (d), we pass to the *natural parametrization* of the exponential family distribution $I|c$, where the natural parameters $\eta = \rho(\theta)$ are a fixed function of the conventional parameters θ. This constraint on the natural parameters ensures that optimization of $L_{\text{cond}}(\eta)$ yields the same answer as optimization of $L_{\text{cond}}(\theta)$. And finally, in line (e) we relax the natural parameter constraint to get the learning objective for a discriminative classifier, where the parameters η are now free to be optimized. A graphical model depiction of this process is shown in Fig. 4.

In summary, starting with a generative classifier with learning objective $L_{\text{gen}}(\theta)$, we complete steps (a) through (e) to arrive at a discriminative classifier with learning objective $L_{\text{dis}}(\eta)$. We refer to this process as a *discriminative relaxation of a generative classifier* and the resulting classifier is a *discriminative counterpart to the generative classifier*.

goose **ostrich** **limousine**

Figure 5: Results of activity maximization on the ImageNet dataset [26]. For a given class c, activity-maximizing inputs are superpositions of various poses of the object, with distinct patches $\mathcal{P}_i$ containing distinct poses $g^*_{\mathcal{P}_i}$, as predicted by Eq. 46. Figure adapted with permission from the authors.

E Derivation of Closed-Form Expression for Activity-Maximizing Images

Results of running activity maximization are shown in Fig. 5 for completeness. Mathematically, we seek the image I that maximizes the score $S(c|I)$ of a specific object class. Using the DRM, we have

$$\begin{aligned}
\max_I S(c^{(L)}|I) &= \max_I \max_{g\in\mathcal{G}} \langle \frac{1}{\sigma^2}\mu(c^{(L)}, g^{(\ell)})|I\rangle \\
&\propto \max_{g\in\mathcal{G}} \max_I \langle \mu(c^{(L)}, g)|I\rangle \\
&= \max_{g\in\mathcal{G}} \max_{I_{\mathcal{P}_1}} \cdots \max_{I_{\mathcal{P}_p}} \langle \mu(c^{(L)}, g)| \sum_{\mathcal{P}_i\in\mathcal{P}} I_{\mathcal{P}_i}\rangle \\
&= \max_{g\in\mathcal{G}} \sum_{\mathcal{P}_i\in\mathcal{P}} \max_{I_{\mathcal{P}_i}} \langle \mu(c^{(L)}, g)|I_{\mathcal{P}_i}\rangle \\
&= \max_{g\in\mathcal{G}} \sum_{\mathcal{P}_i\in\mathcal{P}} \langle \mu(c^{(L)}, g)|I^*_{\mathcal{P}_i}(c^{(L)}, g)\rangle \\
&= \sum_{\mathcal{P}_i\in\mathcal{P}} \langle \mu(c^{(L)}, g)|I^*_{\mathcal{P}_i}(c^{(L)}, g^*_{\mathcal{P}_i}\rangle,
\end{aligned} \tag{45}$$

where $I^*_{\mathcal{P}_i}(c^{(\ell)}, g) \equiv \operatorname{argmax}_{I_{\mathcal{P}_i}} \langle \mu(c^{(\ell)}, g)|I_{\mathcal{P}_i}\rangle$ and $g^*_{\mathcal{P}_i} = g^*(c^{(\ell)}, \mathcal{P}_i) \equiv \operatorname{argmax}_{g\in\mathcal{G}} \langle \mu(c^{(\ell)}, g)|I^*_{\mathcal{P}_i}(c^{(\ell)}, g)\rangle$. In the third line, the image I is decomposed into P patches $I_{\mathcal{P}_i}$ of the same size as I, with all pixels outside of the patch $\mathcal{P}_i$ set to zero. The $\max_{g\in\mathcal{G}}$ operator finds the most probable $g^*_{\mathcal{P}_i}$ within each patch. The solution I^* of the activity maximization is then the sum of the individual activity-maximizing patches

$$I^* \equiv \sum_{\mathcal{P}_i\in\mathcal{P}} I^*_{\mathcal{P}_i}(c^{(\ell)}, g^*_{\mathcal{P}_i}) \propto \sum_{\mathcal{P}_i\in\mathcal{P}} \mu(c^{(\ell)}, g^*_{\mathcal{P}_i}). \tag{46}$$

F From the DRMM to Decision Trees

In this section we show that, like DCNs, Random Decision Forests (RDFs) can also be derived from the DRMM model. Instead of translational and switching nuisances, we will show that an *additive mutation nuisance process* that generates a hierarchy of categories (e.g., evolution of a taxonomy of living organisms) is at the heart of the RDF.

F.1 The Evolutionary Deep Rendering Mixture Model

We define the *Evolutionary DRMM* (E-DRMM) as a DRMM with an evolutionary tree of categories. Samples from the model are generated by starting from the root ancestor template and randomly

mutating the templates. Each child template is an additive "mutation" of its parent, where the specific mutation does not depend on the parent (see Eq.47 below). At the leaves of the tree, a sample is generated by adding Gaussian pixel noise. Like in the DRMM, given $c^{(L)} \sim \mathrm{Cat}(\pi_{c^{(L)}})$ and $g^{(\ell+1)} \sim \mathrm{Cat}(\pi_{g^{(\ell+1)}})$, with $c^{(L)} \in \mathcal{C}^L$ and $g^{(\ell+1)} \in \mathcal{G}^{\ell+1}$ where $\ell = 1, 2, \cdots, L$, the template $\mu_{c^{(L)}g}$ and the image I are rendered as

$$\begin{aligned}\mu_{c^{(L)}g} &= \Lambda_g \mu_{c^{(L)}} \equiv \Lambda_{g^{(1)}} \cdots \Lambda_{g^{(L)}} \cdot \mu_{c^{(L)}} \\ &\equiv \mu_{c^{(L)}} + \alpha_{g^{(L)}} + \cdots + \alpha_{g^{(1)}}, \quad g = \{g^{(\ell)}\}_{\ell=1}^{L} \\ I &\sim \mathcal{N}(\mu_{c^{(L)}g}, \sigma^2 1_D) \in \mathbb{R}^D.\end{aligned}$$

Here, $\Lambda_{g^{(\ell)}}$ has a special structure due to the additive mutation process: $\Lambda_{g^{(\ell)}} = [\mathbf{1} \,|\, \alpha_{g^{(\ell)}}]$, where $\mathbf{1}$ is the identity matrix. The rendering path represents template evolution and is defined as the sequence $(c^{(L)}, g^{(L)}, \ldots, g^{(\ell)}, \ldots, g^{(1)})$ from the root ancestor template down to the individual pixels at $\ell = 0$. $\mu_{c^{(L)}}$ is an abstract template for the root ancestor $c^{(L)}$, and $\sum_\ell \alpha_{g^{(\ell)}}$ represents the sequence of local nuisance transformations, in this case, the accumulation of many additive mutations.

As with the DRMM, we can cast the E-DRMM into an incremental form by defining an intermediate class $c^{(\ell)} \equiv (c^{(L)}, g^{(L)}, \ldots, g^{(\ell+1)})$ that intuitively represents a partial evolutionary path up to level ℓ. Then, the mutation from level $\ell + 1$ to ℓ can be written as

$$\mu_{c^{(\ell)}} = \Lambda_{g^{(\ell+1)}} \cdot \mu_{c^{(\ell+1)}} = \mu_{c^{(\ell+1)}} + \alpha_{g^{(\ell+1)}}. \tag{47}$$

Here, $\alpha_{g^{(\ell)}}$ is the mutation added to the template at level ℓ in the evolutionary tree.

F.2 Inference with the E-DRM Yields a Decision Tree

Since the E-DRMM is an RMM with a hierarchical prior on the rendered templates, we can use Eq.3 to derive the E-DRMM inference algorithm for $\hat{c}^{(L)}(I)$ as:

$$\begin{aligned}\hat{c}^{(L)}(I) &= \operatorname*{argmax}_{c^{(L)} \in \mathcal{C}^L} \max_{g \in \mathcal{G}} \langle \mu_{c^{(L)}} + \alpha_{g^{(L)}} + \cdots + \alpha_{g^{(1)}} | I \rangle \\ &= \operatorname*{argmax}_{c^{(L)} \in \mathcal{C}^L} \max_{g^{(1)} \in \mathcal{G}^1} \cdots \max_{g^{(L-1)} \in \mathcal{G}^{L-1}} \langle \underbrace{\mu_{c^{(L)}} + \alpha_{g^{(L)*}}}_{\equiv \mu_{c^{(L-1)}}} + \cdots + \alpha_{g^{(1)}} | I \rangle \\ &\cdots \\ &\equiv \operatorname*{argmax}_{c^{(L)} \in \mathcal{C}^L} \langle \mu_{c^{(L)}g^*} | I \rangle.\end{aligned} \tag{48}$$

where $\mu_{c^{(\ell)}}$ has been defined in the second line. Here, we assume that the sub-trees are well-separated. In the last lines, we repeatedly use the distributivity of max over sums, resulting in the iteration

$$\begin{aligned}g^*_{c^{(\ell+1)}} &\equiv \operatorname*{argmax}_{g^{(\ell+1)} \in \mathcal{G}^{\ell+1}} \langle \underbrace{\mu_{c^{(\ell+1)}g^{(\ell+1)}}}_{\equiv W^{(\ell+1)}} | I \rangle \\ &\equiv \mathrm{ChooseChild}(\mathrm{Filter}(I)).\end{aligned} \tag{49}$$

Eqs.48 and 49 define a *Decision Tree*. The leaf label histograms at the end of a decision tree plays a similar role as the SoftMax regression layer in DCNs. Applying bagging [5] on decision trees yield a Random Decision Forest (RDF).

G Unifying the Probabilistic and Neural Network Perspectives

H Additional Experimental Results

H.1 Learned Filters and Image Reconstructions

Filters and reconstructed images are shown in Fig. 6.

H.2 Additional Training Results

More results plus comparison to other related work are given in Table 3.

Table 2: Summary of probabilistic and neural network perspectives for DCNs. The DRMM provides a probabilistic interpretation for all of the common elements of DCNs relating to the underlying model, inference algorithm, and learning rules.

Aspect	**Neural Nets Perspective** *(Deep Convolutional Neural Networks)*	**Probabilistic Perspective** *(Deep Rendering Model)*
Model	Weights and biases of filters at a given layer	Partial Rendering at a given abstraction level/scale
	Number of Layers	Number of Abstraction Levels
	Number of Filters in a layer	Number of Clusters/Classes at a given abstraction level
	Implicit in network weights; can be computed by product of weights over all layers or by activity maximization	Category prototypes are finely detailed versions of coarser-scale super-category prototypes. Fine details are modeled with affine nuisance transformations.
Inference	Forward propagation thru DCN	Exact bottom-up inference via Max-Sum Message Passing (with Max-Product for Nuisance Factorization).
	Input and Output Feature Maps	Probabilistic Max-Sum Messages (real-valued functions of variables nodes)
	Template matching at a given layer (convolutional, locally or fully connected)	Local computation at factor node (log-likelihood of measurements)
	Max-Pooling over local pooling region	Max-Marginalization over Latent Translational Nuisance transformations
	Rectified Linear Unit (ReLU). Sparsifies output activations.	Max-Marginalization over Latent Switching state of Renderer. Low prior probability of being ON.
Learning	Stochastic Gradient Descent	Batch Discriminative EM Algorithm with Fine-to-Coarse E step + Gradient M-step. *No coarse-to-fine pass in E-step*.
	N/A	Full EM Algorithm
	Batch-Normalized SGD	Discriminative Approximation to Full EM (assumes Diagonal Pixel Covariance)

Figure 6: (Left) Filters learned from 60,000 unlabeled MNIST samples and (Right) reconstructed images from the Shallow Rendering Mixture Model

I Model Configurations

In our experiments, configurations of the RFM and 2-layer DRFM are similar to LeNet5 [11] and its variants. Also, configurations of the 5-layer DRMM (for MNIST) and the 9-layer DRMM (for CIFAR10) are similar to Conv-Small and Conv-Large architectures in [29, 20], respectively.

Table 3: Test error (%) for supervised, unsupervised and semi-supervised training on MNIST using $N_U = 60K$ unlabeled images and $N_L \in \{100, 600, 1K, 3K, 60K\}$ labeled images.

Model	**Test Error (%)**				
	$N_L = 100$	$N_L = 600$	$N_L = 1K$	$N_L = 3K$	$N_L = 60K$
RFM sup	-	-	-	-	1.21
Convnet 1-layer sup	-	-	-	-	1.30
DRMM 5-layer sup	-	-	-	-	0.89
Convnet 5-layer sup	-	-	-	-	**0.81**
RFM unsup-pretr	16.2	5.65	4.64	2.95	1.17
DRMM 5-layer unsup-pretr	**12.03**	**3.61**	**2.73**	**1.68**	**0.58**
SWWAE unsup-pretr [34]	-	9.80	6.135	4.41	-
RFM semi-sup	14.47	5.61	4.67	2.96	1.27
DRMM 5-layer semi-sup	3.50	**1.56**	1.67	**0.91**	0.51
Convnet [11]	22.98	7.86	6.45	3.35	-
TSVM [33]	16.81	6.16	5.38	3.45	-
CAE [22]	13.47	6.3	4.77	3.22	-
MTC [21]	12.03	5.13	3.64	2.57	-
PL-DAE [12]	10.49	5.03	3.46	2.69	-
WTA-AE [14]	-	2.37	1.92	-	-
SWWAE no dropout [34]	9.17 ± 0.11	4.16 ± 0.11	3.39 ± 0.01	2.50 ± 0.01	-
SWWAE with dropout [34]	8.71 ± 0.34	3.31 ± 0.40	2.83 ± 0.10	2.10 ± 0.22	-
M1+TSVM [8]	11.82 ± 0.25	5.72	4.24	3.49	-
M1+M2 [8]	3.33 ± 0.14	2.59 ± 0.05	2.40 ± 0.02	2.18 ± 0.04	-
Skip Deep Generative Model [13]	1.32	-	-	-	-
LadderNetwork [20]	1.06 ± 0.37	-	$\mathbf{0.84 \pm 0.08}$	-	-
Auxiliary Deep Generative Model [13]	0.96	-	-	-	-
ImprovedGAN [24]	$\mathbf{0.93 \pm 0.065}$	-	-	-	-
catGAN [28]	1.39 ± 0.28	-	-	-	-

Table 4: Test error rates (%) between 2-layer DRMM and 9-layer DRMM trained with semi-supervised EG and other best published results on CIFAR10 using $N_U = 50K$ unlabeled images and $N_L \in \{4K, 50K\}$ labeled images

Model	$N_L = 4K$	$N_L = 50K$
Convnet [11]	43.90	27.17
Conv-Large [29]	-	9.27
CatGAN [28]	19.58 ± 0.46	9.38
ImprovedGAN [24]	$\mathbf{18.63 \pm 2.32}$	-
LadderNetwork [20]	20.40 ± 0.47	-
DRMM 2-layer	39.2	24.60
DRMM 9-layer	23.24	11.37

Table 5: Comparison of RFM, 2-layer DRMM and 5-layer DRMM against Stacked What-Where Auto-encoders with various regularization approaches on the MNIST dataset. N is the number of labeled images used, and there is no extra unlabeled image.

Model	$N = 100$	$N = 600$	$N = 1K$	$N = 3K$
SWWAE (3 layers) [34]	$\mathbf{10.66 \pm 0.55}$	4.35 ± 0.30	3.17 ± 0.17	2.13 ± 0.10
SWWAE (3 layers) + dropout on convolution [34]	14.23 ± 0.94	4.70 ± 0.38	3.37 ± 0.11	2.08 ± 0.10
SWWAE (3 layers) + L1 [34]	10.91 ± 0.29	4.61 ± 0.28	3.55 ± 0.31	2.67 ± 0.25
SWWAE (3 layers) + noL2M [34]	12.41 ± 1.95	4.63 ± 0.24	3.15 ± 0.22	2.08 ± 0.18
Convnet (1 layer)	18.33	10.36	8.07	4.47
RFM (1 layer)	22.68	6.51	4.66	3.55
DRMM 2-layer	12.56	6.50	4.75	2.66
DRMM 5-layer	11.97	**3.70**	**2.72**	**1.60**