

CONSTRAINED STRUCTURED REGRESSION WITH CONVOLUTIONAL NEURAL NETWORKS

Deepak Pathak, Philipp Krähenbühl, Stella X. Yu & Trevor Darrell

Department of Computer Science

UC Berkeley

{pathak, philkr, stellayu, trevor}@berkeley.edu

ABSTRACT

Convolutional Neural Networks (CNNs) have recently emerged as the dominant model in computer vision. If provided with enough training data, they predict almost any visual quantity. In a discrete setting, such as classification, CNNs are not only able to predict a label but often predict a confidence in the form of a probability distribution over the output space. In continuous regression tasks, such a probability estimate is often lacking. We present a regression framework which models the output distribution of neural networks. This output distribution allows us to infer the most likely labeling following a set of physical or modeling constraints. These constraints capture the intricate interplay between different input and output variables, and complement the output of a CNN. However, they may not hold everywhere. Our setup further allows to learn a confidence with which a constraint holds, in the form of a distribution of the constraint satisfaction. We evaluate our approach on the problem of intrinsic image decomposition, and show that constrained structured regression significantly increases the state-of-the-art.

1 INTRODUCTION

Structured regression lies at the heart of some of the most active computer vision problems of our time. Examples include optical flow (Horn & Schunck, 1981), monocular depth estimation (Eigen et al., 2014), intrinsic image decomposition (Barron & Malik, 2015) etc. Convolutional neural networks (CNNs) (Fukushima, 1980; Krizhevsky et al., 2012; LeCun et al., 1989) have greatly advanced the state of the art in all those structured output tasks (Eigen et al., 2014; Narihira et al., 2015a). However CNNs predict each output independently, and thus ignore the intricate interplay of the output variables imposed by physical or modeling constraints. Instead they are forced to learn all physical properties of a scene directly from the training data, and often fail due to the limited capacity of the model.

In this work, we propose to bring those dependencies back to deep structured regression, in the form of constraints on the output space. A constraint ties the output of several regression targets together. In a naive first approach, we learn a standard deep structured regression model and find the closest solution to the predicted structured output that follows the constraints strictly, using a simple Euclidean distance measure. This results in an averaging of the output. It makes only limited use of the training data, and further assumes that the constraints are always satisfied, which is not true in general. For instance, in the task of intrinsic images, Lambert’s law (Barrow & Tenenbaum, 1978) assumes that surfaces have diffused reflectance, which means product of shading and albedo images is equal to the original image. However, this is not true for specular surfaces like mirrors, metals etc (Zhang et al., 1999), such that a naive approach does not perform well in those areas.

To make full use of the training data, we regress not just to a single output variable, but rather a fully factorized distribution over possible outputs. We further predict a distribution over each of our constraints, allowing it not be violated under certain conditions. These distributions capture a the confidence with which the model makes its predictions, or the confidence that a certain constraint holds. A highly confident prediction is peaked around the correct answer, while an uncertain prediction will be more uniform. We use these confidences, and pick the most likely output labeling

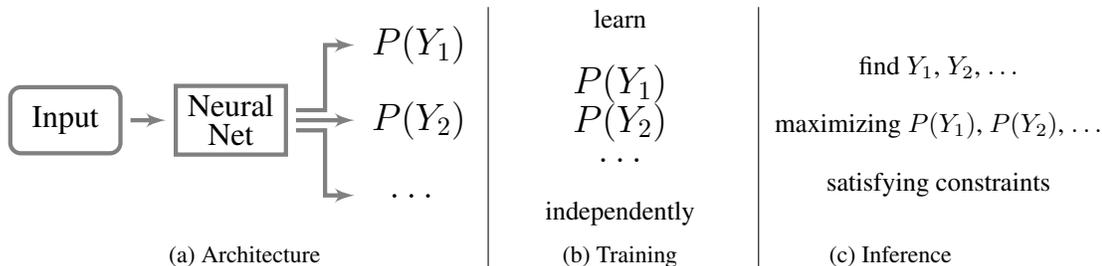


Figure 1: Illustration of our constraint regression framework. We learn the distribution of outputs as well as constraints at training time. At inference, we sample from the learned distribution of output such that the constraints are satisfied.

following our constraints. This allows the model to trust outputs differently during inference, see Figure 1 for an overview of our framework.

We apply our structured regression framework to the problem of intrinsic image decomposition (Barrow & Tenenbaum, 1978). The goal of intrinsic image decomposition is to decompose the input image into albedo (also called reflectance image) and shading images. The output space, in such tasks, has dependencies on the input which can be modeled as physics based constraints such as Lambertian lighting assumption in intrinsic image decomposition (Barrow & Tenenbaum, 1978). At inference time, we find a structured output following those constraints. This alleviates the pressure on the CNN to explicitly learn all physical properties of a scene, and allows it to focus more on the statistical correlations between the input image and the output.

In summary, our constrained regression framework learns not only to capture the ground-truth values, but also to capture the variation or confidence in its own predictions. Moreover, our constraints re-introduce the coupling between albedo and shading that has been ignored in the prior supervised learning. We achieve significant improvement over the state of the art performance and show large visual improvement on MPI Sintel dataset.

2 RELATED WORK

Modeling Weight Uncertainty in Neural Networks In neural networks, uncertainty is usually modeled as distributions over the learned network weights (Denker & Lecun, 1991; MacKay, 1992). If the weights are drawn from a Gaussian distribution then the neural network approximates a Gaussian process in the limit. In such a network, the output distribution is inferred through Bayesian inference. Recently, Blundell et al. (2015) presented a back-propagation based variational technique to learn weight uncertainty. The Bayesian inference based regularization in deep networks is closely related to dropout (Gal & Ghahramani, 2015). Our approach in this paper is not to learn weight uncertainty, but to directly learn the parameters by assuming some output distribution. This allows for much faster inference and training by leveraging the standard feed forward inference and discriminative training for network parameters.

Deep Structured Prediction Most of the problems involve predicting multiple outputs which may be dependent. Such structured prediction tasks are mostly modeled using MRFs. To extend the success of CNN on a small output space to a large one, if the global scoring function on the output can be decomposed into the sum of local scoring functions depending on a small subsets of the output, Chen et al. (2015) shows that the loss function is still representable and there could be an efficient iterative solution. Jaderberg et al. (2015) train CNN to model unaries for the structured prediction task in a max-margin setting for text recognition. Jointly learning CRF and CNN using mean field inference has been shown to achieve great results in segmentation (Zheng et al., 2015). However, most of these works are applied for discrete prediction tasks. For continuous settings, great results have been achieved in depth estimation using CRF formulations, where unary potentials come from a CNN that maps the image to the depth value at single pixels or super-pixels, and where pairwise potentials reflect the desired smoothness of depth predictions between neighboring pixels or segments (Liu et al., 2015; Wang et al., 2015; Li et al., 2015). Compared to these models, not

only our CNN formulation captures the variation in its own predictions, but our constraints also provide an easier and more efficient alternative to higher-order clique potentials. We view structured regression task as constraint satisfaction problem and show that learning variance and distribution of output makes it possible to learn and enforce constraints at inference.

Intrinsic Image Decomposition Intrinsic image decomposition by definition is ambiguous. Traditional approaches seek various physics or statistics based priors on albedo and shading such as sparse and piecewise constant albedo and smooth shading (Horn, 1974; Grosse et al., 2009), or more recently from additional depth images (Lee et al., 2012; Barron & Malik, 2015; Chen & Koltun, 2013). MIT intrinsics, Sintel, and Intrinsic In the Wild datasets (Grosse et al., 2009; Butler et al., 2012; Bell et al., 2014) provide ground-truth for intrinsics and open up new directions with data-driven approaches that have largely got rid of any hand-designed features and complex priors (Narihira et al., 2015b; Zhou et al., 2015; Zoran et al., 2015; Narihira et al., 2015a). Narihira et al. (2015b) regress to a globally consistent ranking of albedo values, however they do not obtain a full intrinsic image decomposition. Zhou et al. (2015) and Zoran et al. (2015) both predict the pairwise ordering of albedo values and then a CRF or a constrained quadratic program to turn this pairwise orderings into a globally consistent and spatially smooth intrinsic image decomposition. Neither method attempted to predict the intrinsic image decomposition directly, due to the lack of training data for real world images. Direct intrinsics (Narihira et al., 2015a) is the first brute-force approach aiming to associate albedo and shading directly with the input image by training convolutional regressors from synthetic ground-truth data. However, lost in the simplicity of regression are the strong coupling constraints in the form of the intrinsics decomposition equation. Our formulation addresses this weakness.

3 PRELIMINARIES

We first introduce the necessary notation and define the common learning objective for structured regression task. We then give a probabilistic interpretation of this regression. Section 4 combines this structured regression with a set of physical inference-time constraints, in our constrained structured regression framework. We learn a complete probabilistic model for structured regression, modeling uncertainties in both the output variables and the constraints. This derivation is general and not application specific. Section 5 then applies our structured regression to intrinsic image decomposition. Section 6 provides more details on our network architecture and training procedure.

Consider the problem of regression to an output $Y \in \mathbb{R}^N$ given an input image I . A deep convolutional regression approximates $Y = f(I; \theta)$ with a multi-layer convolutional neural network with parameters θ . Those parameters θ are learned from a set of training images $\mathcal{I} = \{I_1, I_2, \dots\}$ and corresponding ground truth targets $\tilde{\mathcal{Y}} = \{\tilde{Y}_1, \tilde{Y}_2, \dots\}$. One of the most common learning objectives is a Euclidean loss :

$$\ell(I, \tilde{Y}) = (f(I; \theta) - \tilde{Y})^2, \quad (1)$$

for each image $I \in \mathcal{I}$ and target $\tilde{Y} \in \tilde{\mathcal{Y}}$ in our training set. This Euclidean loss has a nice probabilistic interpretation as the negative log-likelihood of a fully factorized normal distribution $Y_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ with mean $\mu_k = f_k(I; \theta)$ and unit standard deviation $\sigma_k = 1$ for each output Y_k (Koller & Friedman, 2009). In the following section, we use this probabilistic interpretation to enforce constraints on the output space, and find the most likely output Y following these constraints.

4 CONSTRAINED STRUCTURED REGRESSION

Many structured tasks have an inter-dependency of input and output spaces, which may be known either due to physical properties of the problem or due to a modeling choice. Let's suppose that this knowledge is captured by the following family of constraints:

$$\mathcal{G}(Y, I) = 0, \quad (2)$$

where \mathcal{G} is a constraint function acting on the output Y and the input image I .

In practice, we limit ourselves to affine constraints $\mathcal{G}(Y, I) = AY + BI + c$, which results in an efficient inference due to the convexity of constrained region. Note that such a property, although not necessary, is desirable and thus informs our modeling choice. For instance, we perform regression in log domain in intrinsics so that constraints are affine. See Section 5 for more details.

4.1 CONSTRAINED INFERENCE

During inference, we want to find the most likely output estimate Y under the probability distribution predicted by the CNN such that it follows our known family of constraints. Recall from Section 3 that all outputs are modeled by a normal distribution $Y \sim \mathcal{N}(f(I; \theta), 1)$, where the mean is learned. Lets denote the probability of an output Y as $P_{\mathcal{N}}(Y|\theta) = \mathcal{N}(Y; f(I; \theta), 1)$. Thus, finding the most likely output Y can be written as a constrained optimization problem, at inference time, as follows:

$$\begin{aligned} \underset{Y}{\text{minimize}} \quad & -\log P_{\mathcal{N}}(Y|\theta) = (Y - f(I; \theta))^2 \\ \text{subject to} \quad & \mathcal{G}(Y, I) = 0. \end{aligned} \tag{3}$$

If constraints are affine in the output Y , this inference is convex for a fixed set of training parameters. Specifically, this is a quadratic program because the negative log likelihood $-\log P_{\mathcal{N}}$ is a quadratic function, and thus can be solved either in closed form or using QP solvers.

Note that in Equation (3), all outputs Y_k are assumed to have the same distribution with different mean values. This is obviously not true, as some quantities are easier to estimate than others, hence the CNN should be more confident in some areas than others. We will now show how to learn a more general capturing a full distribution for each output Y_k .

4.2 PROBABILISTIC REGRESSION

We use the probabilistic interpretation of the Euclidean loss (1) and also learn the standard deviation of our output as $\sigma_k^2 = g_k(I; \theta)$. This standard deviation can be interpreted as a simple confidence estimate of our prediction. A small standard deviation signifies a confident prediction, while a large standard deviation corresponds to an uncertain prediction. The resulting learning objective is expressed as follows:

$$\begin{aligned} \ell(I, \tilde{Y}) &= -\sum_k \log P_{\mathcal{N}}(\tilde{Y}_k | \mu_k, \sigma_k^2) \\ &= \frac{1}{2} \sum_k \left(\frac{(f_k(I; \theta) - \tilde{Y})^2}{g_k(I; \theta)} + \log g_k(I; \theta) + \log 2\pi \right), \end{aligned} \tag{4}$$

where $P_{\mathcal{N}}$ is a normal distribution, and both f and g are convolutional neural networks parametrized by θ , learned through the same objective. This is standard negative log-likelihood minimization (Koller & Friedman, 2009). For a fixed g this reduces to Euclidean loss. On the other hand, if g is completely free i.e. no regularization, then the objective reduces to an L1 norm with $g_k = |f_k - \tilde{Y}|$ predicting the magnitude of that norm, hence the predicted error. In practice, we parametrize g as the exponential of the network output to ensure positivity. We also add a small l2 regularization to the objective (4).

This probabilistic regression allows us to reason about which outputs to follow more strictly when enforcing our constraints. However all constraints are still strictly enforced. This is fine for setups where it is known that the constraint have to be satisfied, but it is not true for a general scenario as all constraints might not hold for all outputs. Motivated by this, we now consider to learn the uncertainty in constraints as well and then incorporate it in the inference time optimization, similar to Equation (3).

4.3 PROBABILISTIC CONSTRAINTS

In structured prediction tasks, we have the knowledge of constraints that our output should satisfy. If the constraints are not strictly satisfied everywhere, the easiest mechanism would be to learn the distribution over constraint satisfaction in a similar way as we learned the distribution of output in our structured regression setup. More specifically, for i^{th} constraint we write:

$$\mathcal{G}_i(Y, I) \sim \mathcal{H}(0, \sigma_i^2), \tag{5}$$

where \mathcal{H} is a continuous probability distribution. In our experiments, we restrict \mathcal{H} to be a zero-mean Gaussian or Laplace distribution with learned standard deviation, $\sigma_i^2 = h_i(I; \theta)$, where h is a convolutional neural network with same parameters θ as for outputs. For a given output, if the

standard deviation σ_i is low, then our confidence about the i^{th} constraint being satisfied will be high and vice-versa.

To learn this distribution \mathcal{H} , we again follow the standard negative log-likelihood minimization, i.e. minimize $\sum_i (-\log P_{\mathcal{H}}(\mathcal{G}_i|0, \sigma_i^2))$, as mentioned in detail in Section 4.2. We now see how can we incorporate this constraint modeling in the inference framework.

4.4 PROBABILISTIC CONSTRAINED INFERENCE

We conclude by combining the modeled distributions of outputs and constraints in a joint optimization problem to obtain a better estimate of output. Incorporating distributions of output in the optimization is similar to the one described in Equation (3), however, handling the distribution of constraints is not apparently obvious. To address this, interestingly, we can express Equation (5) in terms of a slack variable. This interpretation reduces \mathcal{H} to a distribution over slack on the family of constraints \mathcal{G} . Thus, we find most likely output Y subject to certain constraints such that the slack on the constraints follows distribution \mathcal{H} . This is written as follows:

$$\begin{aligned} & \underset{Y, \epsilon}{\text{minimize}} && -\log P_{\mathcal{N}}(Y|\theta) - \log P_{\mathcal{H}}(\xi|\theta) \\ & \text{subject to} && \mathcal{G}_i(Y, I) = \xi, \end{aligned} \tag{6}$$

where ξ is the slack random variable. The resulting formulation, if constraints are affine, is again convex for a fixed training parameters. For constraints modeled as a Gaussian or Laplacian distribution, the optimization (6) is again a Quadratic Program. Hence, it can be optimized easily for an accurate solution.

We can better understand the Equation (6) in terms of confidences i.e. standard deviation. If constraint has higher confidence, we would like to change the outputs with less confidence more and stick to the predicted value in case of outputs with more confidence to obtain the final estimates, and vice-versa. This process is jointly achieved in this optimization since on taking log-likelihood the standard deviation becomes the weight of the squared error as discussed in Section 3.

4.5 RELATION TO CONDITIONAL RANDOM FIELDS

So far, we have looked at the inference process from a constrained optimization perspective. We want to enforce some known constraints on a structured output space. We first introduce the hard constraint scenario in Equation (3), then learn the distribution over outputs (Eq. 4) and constraints (Eq. 5) to jointly phrase the final optimization in Equation (6) as a constrained optimization problem with a distribution over slack. In this section, we discuss an alternative perspective to directly justify the Equation (6) by showing a connection with conditional random fields (CRF).

Lets ignore the l2 regularization for this analogy. We can relate the distribution defined for our output in Section 3 to the unary potential of a CRF i.e. $\Psi_k(Y_k; \theta) = P_{\mathcal{N}}(Y_k|\theta)$. Further, the constraint distribution to define the joint distribution over outputs can be re-parameterized by the constraint as $\Psi_i(Y; \theta) = P_{\mathcal{H}}(\mathcal{G}_i(Y)|\theta)$. The joint inference of such a CRF would be similar to the Langrangian dual of the problem (6). The log-likelihood learning formulation described in our setup can then be theoretically justified using the piecewise-learning procedure proposed by Sutton & McCallum (2009). During piecewise learning of factors, the contribution of individual outputs often gets over-counted which is determined using cross validation (Shotton et al., 2009; Sutton & McCallum, 2009). However, we empirically found the cross-validation not to be necessary and weight the unary and pairwise terms equally, which gets handled by Langrangian.

5 INTRINSIC IMAGE DECOMPOSITION

As a proof of concept, we apply our approach to the intrinsic image decomposition. Given an input image I , the task is to decompose it into albedo A and shading S (also called reflectance) image. The actual color of the contents in image is captured in albedo, while shading captures the illumination and shape information. The physical constraint which is usually considered in this setting is based on Lambert’s law, i.e. $I = A * S$. However, it holds only for the surfaces where incident light is diffused equally in all directions, also called Lambertian surfaces. It is not true for Specular surfaces,

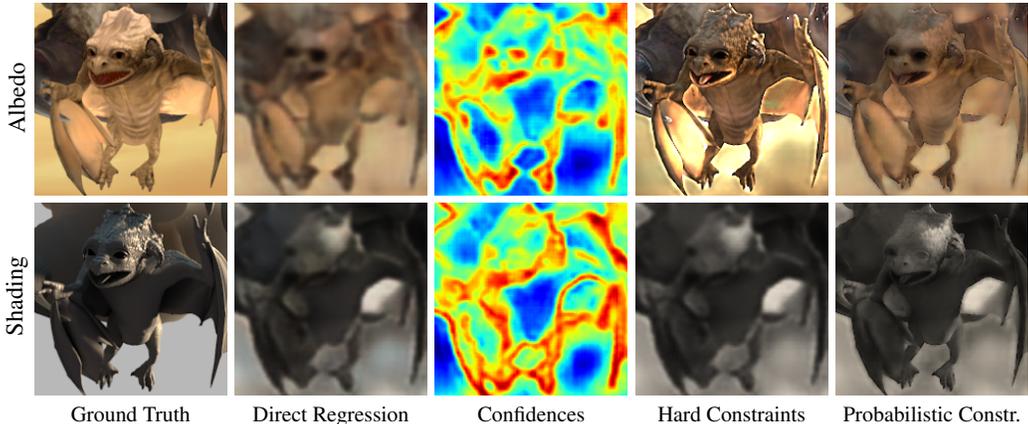


Figure 2: Qualitative ablation for evaluating the contribution of our constrained structured regression pipeline. Direct regression doesn’t capture fine details. If we naively enforce hard constraints, we see color artifacts due to non-lambertian surfaces. Our final probabilistic constrained output results into a detailed output with appropriate albedo.

like mirrors, which maintain the direction of reflection. This constraint provides an important cue for inferring A and S given I . Mostly, this constraint is assumed to be true everywhere and only one of A or S is optimized (Chen & Koltun, 2013; Barron & Malik, 2015). Narihira et al. (2015a) use CNN to regress to the targets A, S , hoping that it would implicitly learn these dependencies which is not true since the outputs are blurry and doesn’t contain any high frequency information. By reasoning and combining evidence from both albedo and shading CNN, we achieve significant improvement.

To keep the optimization convex at inference, we aim to keep the constraints affine in terms of output. In this setting, working in log domain ensures that the intrinsic constraint is affine. Thus our constraint is $G = A + S - I = 0$, where I, A and S are the image, albedo and shading in log domain respectively. The albedo A is a three channeled RGB image. Shading $S = B \cdot C$ is modeled by single channel gray-scale image B , capturing the light intensity, and global color value C , capturing the color of the light. We found this formulation to be numerically more stable than modeling S using three scalars per pixel.

We now discuss the instantiation of our method for this task.

LEARNING

We learn both the distribution of outputs as well as the constraints as discussed in Section 4. Our outputs are $Y_1 = A$ and $Y_2 = S$. Their distribution is learned using the log-likelihood loss defined in Equation (4). We learn the distribution of Lambertian constraint G using loss described in Section 4.3. Ideally one should model \mathcal{H} as Laplacian distribution, as the true violation follows it closely. However in practice we found that Gaussian works equally well and is easier to optimize.

We tie the standard deviation of albedo distribution σ_A across channels at each pixel. Thus, our estimated standard deviations are single channeled heatmaps for each of the output as well as constraint. To summarize, our CNN at training learns to predict three channeled albedo A , and single channeled shading $S, \sigma_A, \sigma_S, \sigma_G$ at every pixel.

INFERENCE

During inference, we estimate the albedo A and shading S using the probabilistic constrained optimization defined in Section 4.4. Special care has to be taken when optimizing the shading S factored into gray scale values B and global color C . While the overall problem is still convex, we found it easier to minimize objective (6) using an alternating minimization. We first keep the current estimate for albedo A and B fixed and optimize for a global light color C , then keep A and C fixed

MPI Sintel	MSE			LMSE			DSSIM		
	Albedo	Shading	Avg	Albedo	Shading	Avg	Albedo	Shading	Avg
Scene Split [Disjoint Training and Testing Scenes]:									
L2 loss	1.93%	2.20%	2.07%	1.22%	1.51%	1.36%	21.22%	15.84%	18.53%
L2 loss + hard constr.	3.96%	2.21%	3.09%	2.22%	1.52%	1.87%	20.39%	15.84%	18.12%
Distr. loss	1.95%	2.07%	2.01%	1.24%	1.43%	1.33%	22.08%	16.44%	19.26%
Distr. loss + hard constr.	3.45%	2.07%	2.76%	2.11%	1.43%	1.77%	19.83%	16.28%	18.06%
Distr. loss + learned constr.	1.84%	1.93%	1.89%	1.15%	1.33%	1.24%	17.78%	13.72%	15.75%

Table 1: Evaluation of different ablations of our model on the MPI Sintel Dataset. We compare our full constrained regression framework with a simple l2 loss with and without hard constraints. Note that all these loss functions are invariant to scale as discussed in Section 6.

and optimize for B , and finally solve for the albedo A keeping the shading $S = B \cdot C$ fixed. Each of those steps has a closed form solution.

6 IMPLEMENTATION DETAILS

We evaluate our constrained structured regression framework on the task of intrinsic image decomposition. Our CNN architecture for modeling the output and constraints distribution is derived from VGG architecture (Simonyan & Zisserman, 2015). It is similar to VGG from conv1 to conv4, which is then followed by three stride 2 deconvolution to upsample the output to original size. Each of these deconvolution layers are followed by rectified linear unit (ReLU) activation function. We initialize truncated-VGG layers using Imagenet (Russakovsky et al., 2015) pretrained weights, and randomly initialize the deconvolution layers.

Notice that albedo and shading for a given image are measured upto a scale, similar to depth. Thus, we use scale invariant learning procedure (Eigen et al., 2014). We adjust the scale of our prediction by a global constant for each of our output and constraint distribution. For log domain, it is equivalent to adjusting optimal shift i.e. for output Y_k with target \tilde{Y}_k , optimal shift α is computed as follows

$$\underset{\alpha}{\text{minimize}} \quad - \|\alpha + Y_k - \tilde{Y}_k\|_2^2 + \beta \|\alpha\|_2^2,$$

where $\beta = 0.5$ is the regularization coefficient. Note that regularization β is crucial to prevent the network from learning arbitrary scaling. We shift output as $Y_k = Y_k + \alpha$ and continue the learning procedure as described in Section 4.

Our complete implementation is done in caffe (Jia et al., 2014) and we use ADAM (Kingma & Ba, 2015) as stochastic gradient solver. The learning rate is kept fixed at $1e - 4$ throughout the process. Our method trains in 5K-10K iterations, taking about 4-5 hours using cpu implementation of constraint structured loss. Complete source code and trained models will be released upon acceptance of publication.

7 EXPERIMENTS

For our evaluation of intrinsic image decomposition, we use MPI Sintel Dataset (Butler et al., 2012). We use the same setup followed by Narihira et al. (2015a) and Chen & Koltun (2013). Dataset contains total of 890 images with albedo and shading ground truth from 18 scenes. Following Narihira et al. (2015a), we do not train on the image-based split proposed by Chen & Koltun (2013) because there is a large overlap in the image content in training and testing set. Instead, we report results on the scene split and report the average error using 2-fold cross-validation, similar to Narihira et al. (2015a).

We compare all algorithm using the mean squared error (MSE), local mean squared error (LMSE) and structural dissimilarity (DSSIM) metric as proposed by Chen & Koltun (2013). While both MSE and LMSE measure the raw pixelwise difference between predictions, DSSIM tried to capture the perceptual visual difference between outputs.

MPI Sintel	MSE			LMSE			DSSIM		
	Albedo	Shading	Avg	Albedo	Shading	Avg	Albedo	Shading	Avg
Image Split [Overlapping Training and Testing Scenes]:									
Grosche et al. (2009)	6.06%	7.27%	6.67%	3.66%	4.19%	3.93%	22.70%	24.00%	23.35%
Lee et al. (2012)	4.63%	5.07%	4.85%	2.24%	1.92%	2.08%	19.90%	17.70%	18.80%
Barron & Malik (2015)	4.20%	4.36%	4.28%	2.98%	2.64%	2.81%	21.00%	20.60%	20.80%
Chen & Koltun (2013)	3.07%	2.77%	2.92%	1.85%	1.90%	1.88%	19.60%	16.50%	18.05%
Scene Split [Disjoint Training and Testing Scenes]:									
Narihira et al. (2015a)	2.09%	2.21%	2.15%	1.35%	1.44%	1.39%	20.81%	16.08%	18.44%
Our Method	1.84%	1.93%	1.89%	1.15%	1.33%	1.24%	17.78%	13.72%	15.75%
Relative Gain	13.59%	14.51%	13.76%	17.39%	8.27%	12.10%	17.04%	17.20%	17.08%

Table 2: Comparison of our constrained regression framework with all prior state of the art on MPI Sintel Dataset. Note that our final model leads to a relative improvement between 10% to 20% throughout all metrics.

Table 1 compares different variations of our algorithm. We start with a baseline L2 loss on our VGG based architecture. While this L2 loss performs quite well, combining it with inference time constraints significantly degrades the performance. Our distribution based loss performs similarly to the simple L2 loss, however it absorbs the constraints much better. Even in this setting constraints actually hurt the raw numeric performance in all settings except for SSIM, meaning that the L2 pixelwise error increased while the perceptual quality of the output slightly increased. However, as soon as we learn the constraint satisfaction our structured regression framework significantly outperforms the baselines. The qualitative ablation depicting different variations is shown in Figure 2.

Finally, we compare our direct intrinsic regression output to the state-of-the-art intrinsic image decompositions in Table 2. Note that we outperform the best prior work by anywhere between 10% and 20% in relative terms. Qualitative comparison of our results is shown in Figure 3.

8 DISCUSSION

We proposed a generic framework for deep structured regression, exploiting inference time constraints. The constrained structured regression framework is general and easy to train. The method is potentially applicable to a broad set of application areas e.g., monocular depth estimation Eigen et al. (2014), optical flow prediction Horn & Schunck (1981), etc. We illustrated performance on a state of the art intrinsic image decomposition task, where our results confirmed that adding structure back to inference can substantially improve the performance of deep regression models both visually and quantitatively.

REFERENCES

- Barron, Jonathan T and Malik, Jitendra. Shape, illumination, and reflectance from shading. *PAMI*, 2015. 1, 3, 6, 8
- Barrow, HG and Tenenbaum, JM. Recovering intrinsic scene characteristics from images. *Computer vision systems*, pp. 2, 1978. 1, 2
- Bell, Sean, Bala, Kavita, and Snavely, Noah. Intrinsic images in the wild. *Siggraph*, 2014. 3
- Blundell, Charles, Cornebise, Julien, Kavukcuoglu, Koray, and Wierstra, Daan. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015. 2
- Butler, Daniel J, Wulff, Jonas, Stanley, Garrett B, and Black, Michael J. A naturalistic open source movie for optical flow evaluation. In *ECCV*, pp. 611–625. Springer, 2012. 3, 7
- Chen, Liang-Chieh, Schwing, Alexander G, Yuille, Alan L, and Urtasun, Raquel. Learning deep structured models. *ICML*, 2015. 2



Figure 3: Qualitative comparison of the results on MPI Sintel Dataset. Note that all approaches, except ours and [Narihira et al. \(2015a\)](#), uses depth image as input. Our approach outperforms all other approaches significantly just using rgb image. Our results are not only correct, but also very detailed contrary to [Narihira et al. \(2015a\)](#).

Chen, Qifeng and Koltun, Vladlen. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, pp. 241–248. IEEE, 2013. [3](#), [6](#), [7](#), [8](#)

Denker, John and Lecun, Yann. Transforming neural-net output levels to probability distributions. In *NIPS*, 1991. [2](#)

Eigen, David, Puhrsch, Christian, and Fergus, Rob. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. [1](#), [7](#), [8](#)

Fukushima, Kunihiko. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980. [1](#)

Gal, Yarin and Ghahramani, Zoubin. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *arXiv:1506.02142*, 2015. [2](#)

Grosse, Roger, Johnson, Micah K, Adelson, Edward H, and Freeman, William T. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009. [3](#), [8](#)

Horn, Berthold K and Schunck, Brian G. Determining optical flow. In *1981 Technical symposium east*, pp. 319–331. International Society for Optics and Photonics, 1981. [1](#), [8](#)

Horn, B.K.P. Determining lightness from an image. *Computer Graphics and Image Processing*, 1974. [3](#)

- Jaderberg, Max, Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Deep structured output learning for unconstrained text recognition. *ICLR*, 2015. 2
- Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, Jonathan, Girshick, Ross B., Guadarrama, Sergio, and Darrell, Trevor. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia, MM*, 2014. 7
- Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *ICLR*, 2015. 7
- Koller, Daphne and Friedman, Nir. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 3, 4
- Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1
- LeCun, Yann, Boser, Bernhard, Denker, John S, Henderson, Donnie, Howard, Richard E, Hubbard, Wayne, and Jackel, Lawrence D. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 1
- Lee, Kyong Joon, Zhao, Qi, Tong, Xin, Gong, Minmin, Izadi, Shahram, Lee, Sang Uk, Tan, Ping, and Lin, Stephen. Estimation of intrinsic image sequences from image+ depth video. In *ECCV*. 2012. 3, 8
- Li, Bo, Shen, Chunhua, Dai, Yuchao, van den Hengel, Anton, and He, Mingyi. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*, June 2015. 2
- Liu, Fayao, Shen, Chunhua, and Lin, Guosheng. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, June 2015. 2
- MacKay, David JC. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. 2
- Narihira, Takuya, Maire, Michael, and Yu, Stella X. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015a. 1, 3, 6, 7, 8, 9
- Narihira, Takuya, Maire, Michael, and Yu, Stella X. Learning lightness from human judgement on relative reflectance. In *CVPR*, 2015b. 3
- Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, Berg, Alexander C., and Fei-Fei, Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 7
- Shotton, Jamie, Winn, John, Rother, Carsten, and Criminisi, Antonio. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009. 5
- Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 7
- Sutton, Charles and McCallum, Andrew. Piecewise training for structured prediction. *Machine learning*, 2009. 5
- Wang, Peng, Shen, Xiaohui, Lin, Zhe, Cohen, Scott, Price, Brian, and Yuille, Alan L. Towards unified depth and semantic prediction from a single image. In *CVPR*, June 2015. 2
- Zhang, Ruo, Tsai, Ping-Sing, Cryer, James Edwin, and Shah, Mubarak. Shape-from-shading: a survey. *PAMI*, 21(8):690–706, 1999. 1
- Zheng, Shuai, Jayasumana, Sadeep, Romera-Paredes, Bernardino, Vineet, Vibhav, Su, Zhizhong, Du, Dalong, Huang, Chang, and Torr, Philip. Conditional random fields as recurrent neural networks. *ICCV*, 2015. 2
- Zhou, Tinghui, Krähenbühl, Philipp, and Efros, Alexei A. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*, 2015. 3
- Zoran, Daniel, Isola, Phillip, Krishnan, Dilip, and Freeman, William T. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015. 3