DPF: Learning Dense Prediction Fields with Weak Supervision

Xiaoxue Chen[1], Yuhang Zheng[2], Yupeng Zheng[3]
Qiang Zhou[1], Hao Zhao[1], Guyue Zhou[1], Ya-Qin Zhang[1]
[1]AIR, Tsinghua University [2]BUAA [3]CASIA
{chenxiaoxue, zhaohao}@air.tsinghua.edu.cn, zyh_021@buaa.edu.cn

Abstract

Nowadays, many visual scene understanding problems are addressed by dense prediction networks. But pixel-wise dense annotations are very expensive (e.g., for scene parsing) or impossible (e.g., for intrinsic image decomposition), motivating us to leverage cheap point-level weak supervision. However, existing pointly-supervised methods still use the same architecture designed for full supervision. In stark contrast to them, we propose a new paradigm that makes predictions for ***point coordinate queries****, as inspired by the recent success of implicit representations, like distance or radiance fields. As such, the method is named as dense prediction fields (DPFs). DPFs generate expressive intermediate features for continuous sub-pixel locations, thus allowing outputs of an arbitrary resolution. DPFs are naturally compatible with point-level supervision. We showcase the effectiveness of DPFs using two substantially different tasks: high-level semantic parsing and low-level intrinsic image decomposition. In these two cases, supervision comes in the form of single-point semantic category and two-point relative reflectance, respectively. As benchmarked by three large-scale public datasets PASCALContext, ADE20K and IIW, DPFs set new state-of-the-art performance on all of them with significant margins. Code can be accessed at https://github.com/cxx226/DPF.*

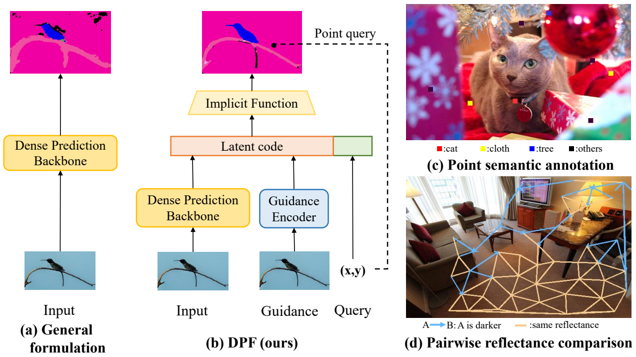

Figure 1. (a) Existing dense prediction formulation. (b) Our DPF formulation. (c) Semantic annotation for single points. (d) Pairwise reflectance annotation between two points.

1. Introduction

The field of visual scene understanding aims to recover various scene properties from input images, e.g., semantic labels [24], depth values [49] [66], edge existence [1] or action affordance [10]. Successful and comprehensive scene understanding is the cornerstone of various emerging artificial intelligence applications, like autonomous driving, intelligent robots or smart manufacturing. Albeit difficult, this field has seen great progress thanks to end-to-end dense prediction networks like DPT [48] and large-scale densely-labelled datasets like ADE20K [67]. If we can densely label every property that we care about, totally solving the visual scene understanding problem seems a matter of time.

However, dense annotations are usually too expensive or impossible to obtain. According to the Cityscapes paper [13], it takes 1.5 hours to generate a high-quality semantic annotation map for a single image. What's worse, for the problem of decomposing an image into reflectance and shading [1], it's impossible for humans to provide pixel-wise ground truth values. As such, the largest intrinsic image decomposition dataset IIW [7] is annotated in the form of pair-wise reflectance comparison between two points. Annotators are guided to judge whether the reflectance of one point is darker than that of another point or not.

Given the importance of dense prediction and the difficulty of obtaining dense annotations, we focus on learning with point-level weak supervision. Fig. 1-c shows an example of point-level semantic scene parsing annotation. The sole red point on the cat is annotated as *cat*, which is much more cheaper than delineating the cat's contours. Fig. 1-d shows human judgement of relative reflectance annotation between every pair of two points. Since the floor has a constant reflectance, point pairs on the floor are annotated with

[1] Intrinsic image decomposition.

the *equal* label. Since the table has a darker reflectance than the floor, pairs between the table point and the floor point are annotated with the *darker* label.

How could we effectively learn dense prediction models from these kinds of point-level weak supervision? To this end, existing pointly-supervised methods leverage unlabelled points using various techniques like online expansion [47], uncertainty mixture [64] [57] or edge guidance [18]. But they all exploit conventional formulations shown in Fig. 1-a, by converting point-level supervision into dense ground truth maps with *padded ignore values*. By contrast, we seek alternative network architectures that are naturally compatible with point-level supervision. Specifically, we take inspiration from the success of neural implicit representations. DeepSDF [46] takes 3D coordinates as input and predicts signed distance values. NeRF [41] takes 5D coordinates as input and predicts radiance/transparency values. Similarly, our method takes 2D coordinates as input and predicts semantic label or reflectance values, as shown in Fig. 1-b. An intriguing feature of this new scheme is that high-resolution images can be encoded as guidance in a natural way, because this new continuous formulation allows outputs of arbitrarily large or small resolution. Borrowing names from the research community of distance or radiance fields, our method is called **dense prediction fields (DPFs)**.

In order to show that DPF is a strong and generic method, we use two pointly-supervised tasks: semantic scene parsing and intrinsic image decomposition. These two tasks differ in many aspects: (1) Scene parsing is a high-level cognitive understanding task while intrinsic decomposition is a low-level physical understanding task; (2) Scene parsing outputs discrete probability vectors while intrinsic decomposition outputs continuous reflectance/shading values; (3) Scene parsing is annotated with single points while intrinsic decomposition is annotated with two-point pairs. Interestingly and surprisingly, our method achieves new state-of-the-art results on both of them, as benchmarked by three widely used datasets PASCALContext, ADE20K and IIW.

To summarize, the contributions of our work include:

- We propose a novel methodology for learning dense prediction models from point-level weak supervision, named DPF. DPF takes 2D coordinates as inputs and allows outputs of an arbitrary resolution.

- We set new state-of-the-art performance on PASCAL-Context and ADE20K datasets for scene parsing and IIW dataset for intrinsic decomposition with point-level weak supervision. Codes are publicly available.

- With systematic ablations, visualization and analysis, we delve into the mechanism of DPF and reveal that its superior performance is credited to locally smooth embeddings and high-resolution guidance.

2. Related Work

2.1. Intrinsic image decomposition

Complete scene de-rendering [30,33] is a long-term goal in visual intelligence, requiring many properties to be understood, like geometry [29,61], room layout [11,20,27,63], lighting [21,55], and material [16,39,65]. **Intrinsic decomposition** is the minimal formulation that decomposes a natural image into reflectance and shading. Since the problem is severely ill-posed, conventional methods [9, 23, 51–53] resort to optimization algorithms with hand-crafted priors. Recently, many deep learning methods [3–5, 14, 15, 19, 28, 44,71] have been proposed to solve it. [37,38,62] explore to address this problem in unsupervised manners. [43, 69] apply a CNN network to directly predict reflectance or shading. [45] develops a joint bilateral filtering method to leverage strong prior knowledge about reflectance constancy. [18] adopts a guided, edge-preserving domain filter to generate realistic reflectance. [34] proposes a new end-to-end training pipeline that learns better decomposition by leveraging a large-scale synthetic dataset CGIntrinsics. [33, 68] introduce novel lighting representations to obtain a complete scene reconstruction including reflectance, shape, and lighting. IRISFormer [70] adopts a transformer architecture to simultaneously estimate depths, normals, spatially-varying albedo, roughness and lighting from a single image. In this work, we focus on pointly-supervised intrinsic decomposition. Specifically, we benchmark on the IIW dataset [7], which is annotated with sparse, pairwise comparison labels. Although many of the above works are also evaluated on IIW, none of them are specifically designed for point supervision. Instead, our DPF method is **naturally compatible** with point supervision and achieves superior performance compared with all prior works.

2.2. Scene parsing and weak supervision

The goal of scene parsing is to classify all pixels in the image into corresponding categories. However, dense annotation for images, which costs a lot, is still critical to the success of scene parsing. This fact gives rise to the research of dense prediction with weak supervision. One line of works focuses on the usage of pseudo labels. Although prior methods of harvesting pseudo labels are designed in various manners, they rely on proper thresholds [6, 47, 58, 59, 64]. Among all, uncertainty mixture [64] that has the capacity of choosing the threshold automatically achieves strong results on the PASCALContext and ADE20K dataset. Recently, transformer based models have made great progress in scene parsing. The vision transformer (ViT) backbone [17] significantly benefits dense prediction due to its characteristics of maintaining a representation with constant spatial resolution throughout all processing stages and having a global receptive field at every stage. Our method is based on the ViT backbone, leveraging the self-attention mechanism to better propagate supervision signals from sparse points to **all patch tokens**.

2.3. Implicit neural representation

Implicit neural representation is a paradigm that maps coordinates to signals in a specific domain with neural networks. On account of the continuous and differentiable deep implicit function, it can capture intricate details, bringing conspicuous performance in 3D reconstruction [22, 40, 46]. Recent researches also show the effectiveness of the implicit representation on 2D tasks. The Local Implicit Image Function [12] learns a continuous image representation that can be queried in arbitrary resolution. The Joint Implicit Image Function (JIIF) [56] formulates guided depth super-resolution as a neural implicit image interpolation problem. SIREN [54] leverages periodic activation functions for implicit neural representations, which are ideally suited for representing complex natural signals and their derivatives. The Implicit Feature Aliment Function [26] implicitly aligns the feature maps at different levels and is capable of producing segmentation maps in arbitrary resolutions. Based on the fact that point queries and point supervision are inherently compatible, we explore the employment of neural implicit image interpolation with point queries under weak supervision.

3. Method

3.1. Dense Prediction with Point Supervision

Given an input image, dense prediction is the task of predicting an entity of interest (a label or a real number) for each pixel in the image. Previous works [48] [10] usually use pixel-wise annotations as the supervision to train dense prediction models. However, it is time-consuming to annotate in a pixel-wise manner. Sometimes it's **even impossible** to annotate the pixel with a certain value, for example, annotating an in-the-wild image with specific reflectance. Therefore, in this work, we focus on dense prediction with point supervision and propose a novel neural network to resolve it. Specifically, we introduce a dense prediction field (DPF) that predicts a corresponding value for each continuous 2D point on the imaging plane. Moreover, inspired by the recent success of implicit representations [46] [32], we use an implicit neural function to implement the DPFs. Mathematically, given a coordinate query x in the image,

$$\mathrm{DPF}(\mathrm{x}) = \mathrm{v_x}, \quad \mathrm{v_x} \in \mathbb{R}^{\mathrm{c}}, \tag{1}$$

where c is the dimension of predicted entity. Due to its continuous nature, DPF is spatially consistent thus can achieve superior performance under sparse point supervision.

To verify the effectiveness of the proposed dense prediction field, we benchmark it on two different types of pointly-supervised datasets: (1) datasets with sparsely labeled semantic category information like PASCALContext and ADE20K, and (2) datasets labeled with sparse pairwise comparisons like IIW. Both PASCALContext and ADE20K are designed for semantic parsing. While IIW is aimed at decomposing natural images into intrinsic reflectance and shading. These two types of datasets involve different prediction targets and different losses. Experiments show that we achieve SOTA performance on all three datasets which demonstrates the **generalizability** of our method.

3.2. Network Architecture

As depicted in Fig. 2, our network is composed of three components: a dense prediction backbone h_λ, a guidance encoder g_η, and an implicit dense prediction field f_θ. The overall formulation of DPF is:

$$\mathrm{v_x} = \mathrm{f}_\theta(\mathrm{z}, \mathrm{g}, \mathrm{x}), \tag{2}$$

where z and g are the latent codes extracted from h_λ and g_η respectively and x is the point query coordinate.

Dense prediction backbone. Previous works [56] typically formulate an image-based implicit field into:

$$\mathrm{v_x} = \mathrm{f}_\theta(\mathrm{E}(\mathrm{I}), \mathrm{x}) \tag{3}$$

where E is an encoder network to extract low-level visual features as latent code. However, considering the importance of high-level semantic information extracted by specially designed dense prediction networks, we propose a novel paradigm that combines a dense prediction backbone and an implicit field. Specifically, given an input image I, we first feed the image into the dense prediction backbone:

$$\mathrm{V}, \mathrm{z} = \mathrm{h}_\lambda\left(\mathrm{I}\right), \quad \mathrm{V} = \mathrm{h}_{\lambda 1}(\mathrm{z}) \tag{4}$$

$\mathrm{h}_{\lambda 1}$ is the prediction head of h_λ, V is the baseline dense prediction value and z is the high-level features extracted from the output of the intermediate layer of h_λ, specifically z is the output of the penultimate layer (before mapping the number of feature channels to the dimensionality of prediction targets). We impose a loss on V as **auxiliary supervision**, which provides constraints on the predicted value (v in the later Eq. 6) of the implicit field while facilitating the latent code z to acquire corresponding high-level information. This design is ablated in Tab. 5.

The paradigm in Eq. 2 can be applied on top of any plug-and-play dense prediction models. To verify this, we choose a CNN-based network FastFCN [60] and a ViT-based network DPT [48] as the backbones.

Guidance encoder. Guided image filtering [25] is an effective edge-preserving smoothing operator based on a

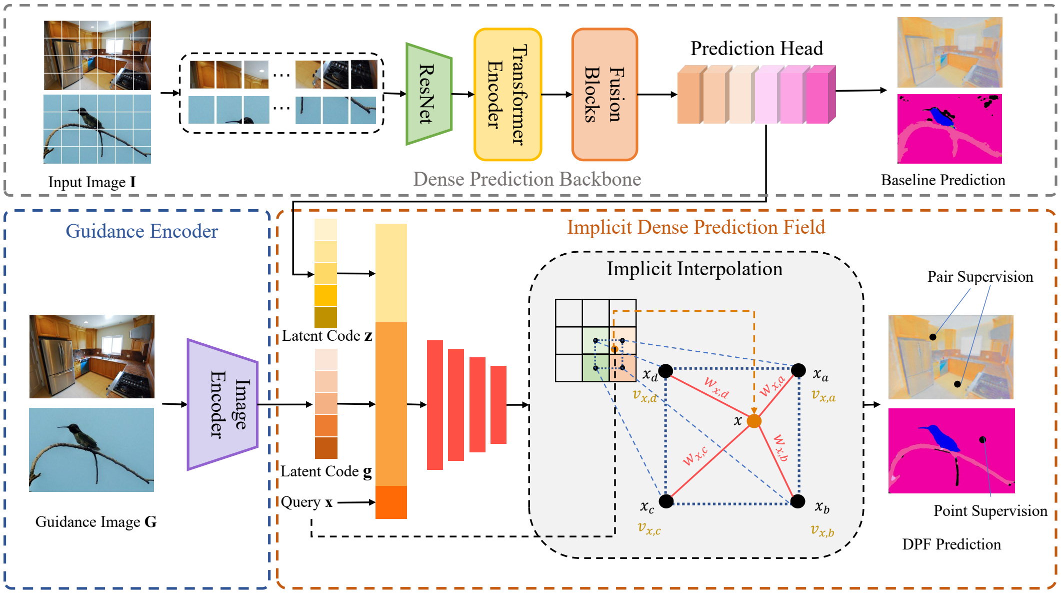

Figure 2. **Overall architecture**. Our model consists of three components: a dense prediction backbone to extract high-level features and make baseline predictions, a guidance encoder to encode guidance features, and an implicit dense prediction field to make predictions at point coordinate queries. The upper figures are for intrinsic decomposition while the lower figures are for scene parsing.

guidance image. Following previous works [18], we also introduce an extra guidance image G. We believe the content of the guidance image can benefit the learning of interpolation parameters (described in Sec 3.3) and make the DPF outputs better aligned with the high-resolution guidance image. We directly use the input image of different resolutions as the guidance image instead of introducing a task-specific guidance map (e.g., the edge guidance in [15, 18]) that requires domain-specific pre-processing.

We use an EDSR [36] network as the guidance encoder, and extract features from the guidance image:

$$\mathrm{g} = \mathrm{g}_{\eta}\left(\mathrm{G}\right) \tag{5}$$

g also serves as a latent code, but it contains low-level local features which are complementary to z. The EDSR model consists of 16 residual blocks without upsampling layers, and we use the output of the last block as g. Both the two latent codes provide important information to support the learning of DPFs. Their effects and differences will be shown in Fig. 7. In the following section, we will describe our implicit dense prediction field in detail.

3.3. Implicit Dense Prediction Field

Given the coordinate x of a point on the image plane, we are aiming to query its value v_x in the dense prediction field. Notably, x can be a random coordinate value sampled from a continuous space, so we can't directly extract the corresponding value from a discrete dense prediction map. A straightforward way to get v_x is to interpolate the dense prediction values of neighbor pixels, as illustrated in Fig. 2 (Implicit Interpolation). Specifically, the corresponding dense prediction value v_x is defined as:

$$v_x = \sum_{i \in N_x} w_{x,i} v_{x,i}, \quad \sum_{i \in N_x} w_{x,i} = 1 \tag{6}$$

where N_x is the set of neighbor pixels of x, $v_{x,i}$ is the dense prediction value of pixel i, $w_{x,i}$ is the interpolation weight between x and i. For the scene parsing tasks with multiple semantic categories, the values are vectors of length c, where c is the number of categories. For the reflectance prediction, the values are scalars. In practice, all the coordinates are normalized into $(-1, 1)$ with the image center as the origin. This normalization step allows us to conveniently combine latent codes of different resolutions (g and z specifically).

Inspired by deep implicit function methods [12, 46, 56], we use a deep neural network to get the interpolation weights and dense prediction values. Given the input image feature z and the guidance feature g, we leverage an MLP to learn the interpolation weights and values between coordinate x and its neighbor pixel i:

$$\hat{w}_{x,i}, v_{x,i} = \mathrm{MLP}(\mathrm{z_i}, \mathrm{g_i}, \gamma(\Delta \mathrm{x})) \tag{7}$$

where $\mathrm{z_i}, \mathrm{g_i}$ is the corresponding latent code of pixel i that is extracted from z and g. $\Delta \mathrm{x} = \mathrm{x_i} - \mathrm{x}$ is a relative coordinate, and $\mathrm{x_i}$ is the coordinate of i. This relative coordinate indicates the spatial affinity between query point x and its neighbor pixel i. Furthermore, we also apply a positional encoding $\gamma(\cdot)$ following [41] to leverage higher frequency spatial signals:

$$\gamma(x) = (sin(2^0 \pi x), cos(2^0 \pi x), ..., sin(2^l \pi x), cos(2^l \pi x)) \tag{8}$$

In practice, we set $l = 9$. After Eq. 7, the interpolation weights are normalized through a softmax layer:

$$w_{x,i} = \frac{\exp(\hat{w}_{x,i})}{\sum_{j \in N_x} \exp(\hat{w}_{x,j})} \tag{9}$$

By integrating the interpolation (Eq. 6) and the calculation of weights and values (Eq. 7, 9), the formulation of our implicit dense prediction field can be represented as:

$$v_x = \mathrm{f}_\theta(\mathrm{z}, \mathrm{g}, \mathrm{x}) \tag{10}$$

where θ is the network parameters.

3.4. Training Loss

To get the prediction of DPFs, we use the coordinate of every pixel in the guide image as queries, and generate a prediction map of the same resolution of the guide image. For the scene parsing task, the number of channels of the prediction map is c, where c is the number of semantic categories. As for intrinsic decomposition, the number of channels is 1. We supervise both the predictions of the dense prediction backbone and DPF using the same kind of loss functions. Specifically, we use a c-way cross-entropy loss for scene parsing datasets.

For the pairwise comparison dataset IIW, there are no absolute ground truth labels available. Instead, given the k-th pair of comparison points $\{k_1, k_2\}$, the relative reflectance annotation J_k is classified into three labels:

$$J_k = \begin{cases} 1 & \text{if } \mathrm{k_1} \text{ is } \textit{darker} \text{ than } \mathrm{k_2}, \\ 2 & \text{if } \mathrm{k_1} \text{ is } \textit{lighter} \text{ than } \mathrm{k_2}, \\ E & \text{if reflectance of } \mathrm{k_1} \text{ and } \mathrm{k_2} \text{ are } \textit{equal}. \end{cases} \tag{11}$$

We denote the predicted reflectance of point k_1 and k_2 as R_{k_1} and R_{k_2}, respectively. We use a standard SVM hinge loss to supervise the pairwise comparison data:

$$\mathcal{L}_k = \begin{cases} max(0, \frac{R_{k_1}}{R_{k_2}} - \frac{1}{1+\delta+\epsilon}) & \text{if } \mathrm{J_k} = 1, \\ max(0, 1 + \delta + \epsilon - \frac{R_{k_1}}{R_{k_2}}) & \text{if } \mathrm{J_k} = 2, \\ max(0, \begin{cases} \frac{1}{1+\delta-\epsilon} - \frac{R_{k_1}}{R_{k_2}}, \\ \frac{R_{k_1}}{R_{k_2}} - (1 + \delta - \epsilon) \end{cases}) & \text{if } \mathrm{J_k} = \mathrm{E}. \end{cases} \tag{12}$$

ϵ and δ are hyper-parameters, and we set $\epsilon = 0.08$ and $\delta = 0.12$ during training.

The total loss for all comparison pairs is defined as:

$$\mathcal{L}_{\text{pairs}} = \sum_{k \in \mathrm{P}} s_k \cdot \mathcal{L}_k \tag{13}$$

where P is the index set of all comparison pairs, and s_k is the confidence score of each annotation provided by the dataset.

4. Experiment

4.1. Datasets and Evaluation Protocols

Intrinsic decomposition. We report results on the IIW dataset [7]. The IIW dataset contains 5,230 indoor scene images, and 872,151 relative reflectance comparison pairs in total. Following the setting of [18], we sort the IIW dataset by image ID, and put the first of every five images into the test set, and the rest into the training set. We employ weighted human disagreement rate (WHDR) as the evaluation metric. The classification of predicted reflectance comparison pairs can be calculated as:

$$\hat{J}_k = \begin{cases} 1 & \text{if } \frac{\mathrm{R_{k2}}}{\mathrm{R_{k1}}} > 1 + \delta, \\ 2 & \text{if } \frac{\mathrm{R_{k1}}}{\mathrm{R_{k2}}} > 1 + \delta, \\ E & otherwise. \end{cases} \tag{14}$$

where δ is the threshold to filter out negligible relative difference, which we set as 0.1 in the evaluation. The WHDR is the error rate of $\hat{J}_k$ when compared with J_k.

Scene parsing. We benchmark DPFs on two scene parsing datasets: PASCALContext [42] and ADE20K [67]. For PASCALContext, 4998 samples are used for training and 5105 samples are used for testing. For ADE20K, 20210 images are used for training and 2000 images are used for testing. PASCALContext has 60 different semantic labels, and ADE20K has 150 different semantic labels. For fair comparison, we use the same point annotations as [64] uses. We choose the mean intersection over union (mIoU) score as the evaluation metric for both datasets.

4.2. Comparisons with SOTA methods

Intrinsic decomposition. We provide the quantitative results of DPF (ViT based) on IIW in Tab. 1. Our model outperforms the previous state-of-the-arts. Specifically, we achieve a 0.1% boost over IRISformer [70], which introduces the OpenRooms (OR) dataset [35] during training. Notably, many methods in Tab.1 introduce synthetic datasets with full intrinsics ground truth, while our method is only trained on the pointly-annotated IIW dataset. Compared with the previous SOTA [18] trained only on IIW, our method promotes WHDR by 2.6%, suggesting the effectiveness of our formulation using pairwise point comparison data. Fig. 3 presents qualitative comparisons on three

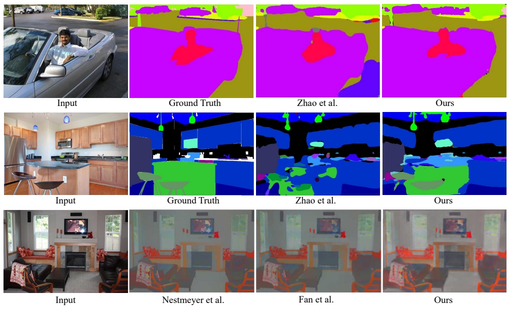

Figure 3. Qualitative comparisons on the PASCALContext (first row), ADE20K (second row) and IIW (last row), respectively.

Method	Training set	WHDR (%) ↓
Sengupta et al. [50]	CGP+IIW	16.8
Li and Snavely [34]	CGI+IIW*	16.2
Li et al. [33]	CGM+IIW	15.9
Zhu et al. [70]	OR+IIW	12.0
Bell et al. [7]	IIW	20.6
Nestmeyer et al. [45]	IIW	17.7
Bi et al. [8]	IIW	17.7
Fan et al. [18]	IIW	14.5
Ours	IIW	**11.9** (+2.6)

Table 1. Quantitative results on IIW. Lower WHDR is better. IIW* indicates augmented IIW comparisons. CGI [34], CGM [33], CGP [50], OR [35] are all intrinsic decomposition datasets with dense labels.

Method	mIoU (%) ↑
Qian et al. [47] w/o Online Ext	29.70
Qian et al. [47] w/ Online Ext	30.00
Zhao et al. [64] w/o rGMM	33.54
Zhao et al. [64] w/ rGMM	36.07
Ours	**45.31** (+9.2)

Table 2. Quantitative results on PASCALContext.

Method	mIoU (%) ↑
Qian et al. [47] w/o Online Ext	19.00
Qian et al. [47] w/ Online Ext	19.60
Zhao et al. [64] w/o rGMM	26.33
Zhao et al. [64] w/ rGMM	28.79
Ours	**33.84** (+5.0)

Table 3. Quantitative results on ADE20K.

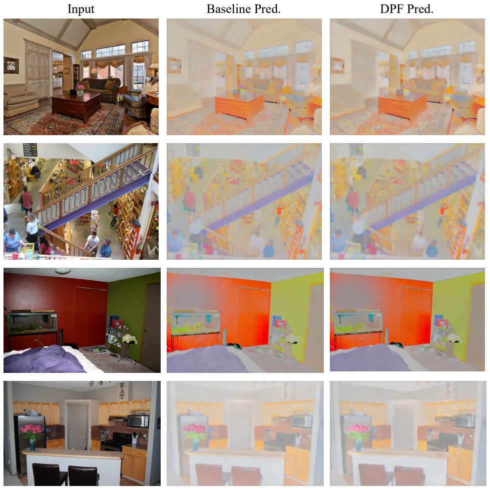

Figure 4. Qualitative prediction results on IIW.

datasets, demonstrating the superior performance of DPFs compared with prior works.

Scene parsing. Tab. 2,3 provide the performance of ViT-based DPFs on PASCALContext and ADE20K respectively. For PASCALContext, the mIoU is significantly promoted from 36.1% to 45.3%. For ADE20K, the mIoU utperforms the previous SOTA by 5.0%. This shows that our model also performs well under the supervision of single sparse labels. On the one hand, this is credited to the attention-based backbone (DPT) we use, which has already shown strong performance in the field of dense prediction tasks due to its global receptive field; on the other hand, our proposed DPF refines the dense prediction results with an implicit neural representation, naturally enabling smoother results under point supervision.

4.3. Effectiveness of DPF on different backbones

To further prove the effectivness of DPF, we train the CNN baseline (FastFCN), ViT baseline (DPT), and DPF with different backbones on all three datasets, and the results are shown in Tab. 4. On all datasets, DPFs outperform the baselines significantly. Specifically, for ViT-based DPF, the mIoU is increased by 4.9% on PASCALContext, performance on ADE20K increases by 3.4%, and WHDR of IIW is decreased by 2.1%, which indicate that DPF conclusively improves pointly-supervised dense prediction. This is credited to the representation of the latent codes, which combines high-level image features, low-level guidance features, and spatial information from relative query coordinates. Meanwhile, the implicit interpolation weights the values of neighbor pixels adaptively, making the dense prediction results more consistent. In addition, using a transformer backbone leads to larger performance improvement than the CNN backbone due to the self-attention mechanism, as it can naturally help the propagation of sparse supervision with global patch interaction.

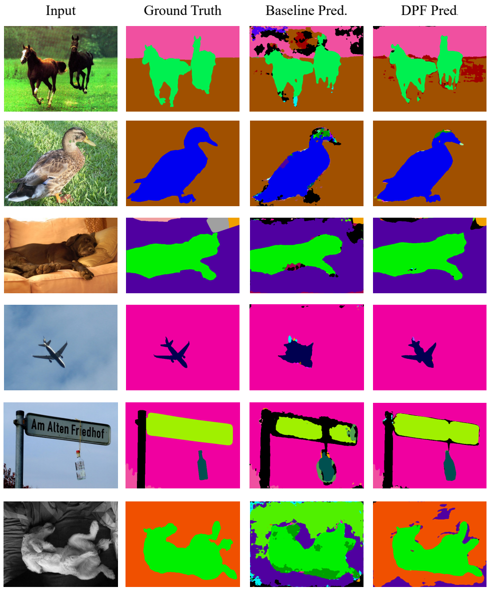

Figure 5. Qualitative prediction results on Pascal. Different colors represent different semantic categories.

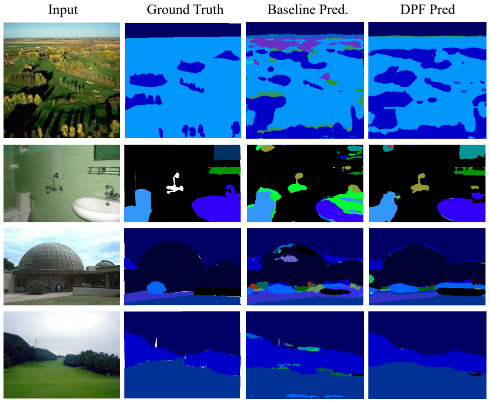

Figure 6. Qualitative prediction results on ADE20K.

Weight visualization. Fig. 8 provides the visualization of the learned interpolation weights. The query pixel is in red, and the four corner pixels' color indicates the learned interpolation weights. Higher weights are in bluer color, while lower weights are greener. It shows that DPF can successfully learn the interpolation weights depending on the location of the query point. When the query point shares the same reflectance or semantic label with its neighbor pixel, the weights will be higher. Conversely, the weights will be lower. We note that this kind of interpretable weight are learned through sparse annotation. This makes the DPF's prediction smoother and more accurate, while respecting the edges in input images.

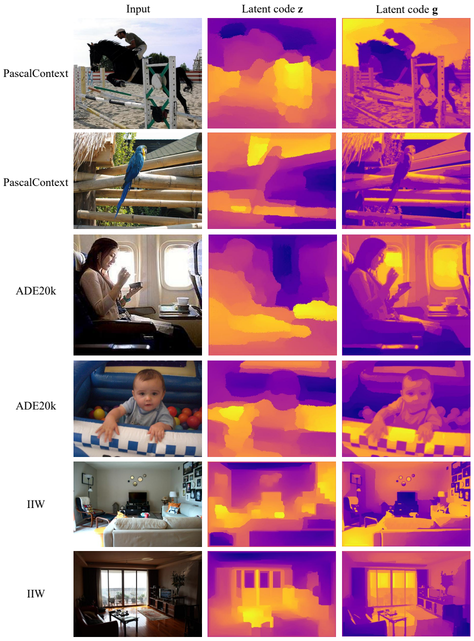

Figure 7. Visualization of t-SNE of latent codes.

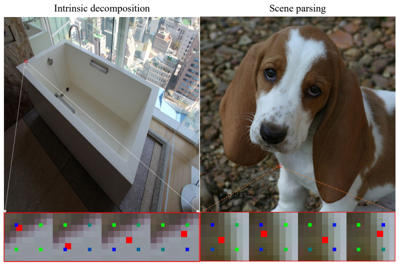

Figure 8. Visualization of the learned interpolation weights.

Qualitative results. We provide visualization results on three datasets in Fig. 4, 5 and 6, respectively. Fig. 5, 6 provides the results on scene parsing. Compared with the baseline prediction, DPF produces more accurate results. Specifically, as shown in the figures, there are a lot of noise patches in the baseline predictions, and these patches are misclassified, making the visualization results look very cluttered. Different from baseline, there are fewer misclassified patches in the DPF predictions. Besides, the predictions on the edge of objects are smoother for DPF, like the edge of the road sign on the 5th row in Fig. 5. Meanwhile, the segmentation of objects is also more precise. Take the airplane on the 4th row (Fig.5) for an example, the shape of the airplane in DPF predictions is more reasonable, while the baseline result is relatively blurry. Fig. 4 presents qualitative results on IIW. As shown in the image, the prediction of DPF is smoother compared with the baseline prediction. For the purple quilt in the third row of Fig. 4, DPF can distinguish reflections and wrinkles, and decompose the quilt

Datasets	PASCAL	ADE20K	IIW
CNN baseline	37.3	26.0	17.9
DPF (CNN)	**38.7** (+1.4)	**27.2** (+1.2)	**17.2** (+0.7)
ViT Baseline	40.4	30.4	14.0
DPF (ViT)	**45.3** (+4.9)	**33.8** (+3.4)	**11.9** (+2.1)

Table 4. DPF Quantitative results using different backbones.

Datasets	**w/o auxiliary**	**w/o guide**	**All**
PASCAL	40.0 (-5.3)	44.5 (-0.8)	**45.3**
ADE20K	28.3 (-5.5)	32.9 (-0.9)	**33.8**
IIW	25.2 (-13.3)	12.5 (-0.6)	**11.9**

Table 5. Quantitative results on the effect of auxiliary supervision and guidance image.

into the same reflectance, while baseline prediction is not as flattened as ours. These results illustrate the capability of DPF on intrinsic decomposition.

4.4. Experiments on Network Architecture

Auxiliary supervision. Tab. 5 investigates the effects of auxiliary supervision on the dense prediction backbone. Removing auxiliary supervision leads to large performance drops on all three datasets. This fact demonstrates that the losses of dense prediction backbone serve as critical supervision roles during the training process. It further verifies that the supervision of V also constrains the predicted values of the implicit field, helping DPFs learn reasonable prediction results. In addition, auxiliary supervision on baseline prediction also benefits latent code z to learn high-level visual features.

Dataset	**Input** **Guide**	**128**	**256**	**512**
PASCAL	/	13.7	27.2	40.4
	128	31.1	-	-
	256	31.8	42.3	-
	512	**32.3**	**42.6**	**45.3**
ADE20K	/	8.9	21.5	30.4
	128	15.9	-	-
	256	16.4	29.0	-
	512	**17.1**	**29.2**	**33.8**
IIW	/	22.2	17.9	14.0
	128	21.4	-	-
	256	21.0	16.5	-
	512	**20.6**	**15.3**	**11.9**

Table 6. Quantitative results with different input image resolutions and guidance resolutions.

Effects of guidance image. We conduct experiments to explore the effect of guidance images. Specifically, we train DPF models without guidance encoder and guidance latent code g on three datasets. The formulation of this simplified DPF is represented as:

$$v_x = f_\theta(z, x) \tag{15}$$

and the results are shown in Tab. 5. It's clear that DPFs with guidance get superior performance. Specifically, for PASCALContext, the mIoU of semantic segmentation is increased by 1.2%, ADE20K performance increases by 1.3%, and WHDR of IIW is decreased by 0.6%. This indicates that guidance images can benefit the learning of interpolation parameters. We believe this plays a similar role to the guidance image in the guided image filter, helping the learning of interpolation parameters, which makes the dense prediction results more consistent. Besides, CRFs like [2, 31] are conventional techniques that work in the same spirit and we provide a comparison in the supplementary.

Resolution of guidance image. We also conduct experiments on the resolutions of guidance images, and the results are presented in Tab. 6. As shown in the table, the results of the DPF with guidance always outperform the baseline model. Specifically, while the mIoU of baseline on Pascal dropped a lot with a 128×128 input image, the DPF with 512×512 guidance image improves the performance by 18.6%. Meanwhile, when the resolution of the input image is the same, the larger the resolution of the guidance image, the better the performance of the model, which has been verified on all three datasets. This further illustrates the importance of guidance, while providing an appealing paradigm that trains low-resolution inputs with high-resolution guidance images.

Visualization of latent code. Fig. 7 presents visualizations of latent codes on three datasets. We use t-SNE to reduce the dimension of latent codes g and z to one and visualize them, respectively. As illustrated in the figure, latent code z encodes high-level features with semantic information, while latent code g focuses on low-level features with clear boundaries. Furthermore, latent code g preserves the details of the original image, but is relatively smoother, which benefits the learning of consistent DPFs.

5. Conclusion

In this paper, we propose dense prediction fields (DPFs), a new paradigm that makes dense value predictions for point coordinate queries. We use an implicit neural function to model the DPFs, which are compatible with point-level supervision. We verify the effectiveness of DPFs using two different tasks: semantic parsing and intrinsic image decomposition. We benchmark DPFs on three datasets including PASCALContext, ADE20K and IIW, and achieve state-of-the-art performance on all three datasets.

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 1

[2] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 524–540. Springer, 2016. 8

[3] Anil S Baslamisli, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Shadingnet: image intrinsics by fine-grained shading decomposition. *International Journal of Computer Vision*, 129(8):2445–2473, 2021. 2

[4] Anil S Baslamisli, Thomas T Groenestege, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Joint learning of intrinsic images and semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–302, 2018. 2

[5] Anil S Baslamisli, Hoang-An Le, and Theo Gevers. Cnn based learning using reflection and retinex models for intrinsic image decomposition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6674–6683, 2018. 2

[6] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 2

[7] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 1, 2, 5, 6

[8] Sai Bi, Xiaoguang Han, and Yizhou Yu. An l 1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Transactions on Graphics (TOG)*, 34(4):1–12, 2015. 6

[9] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *Proceedings of the IEEE international conference on computer vision*, pages 241–248, 2013. 2

[10] Xiaoxue Chen, Tianyu Liu, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Cerberus transformer: Joint semantic, affordance and attribute parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19649–19658, 2022. 1, 3

[11] Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Pq-transformer: Jointly parsing 3d objects and layouts from point clouds. *IEEE Robotics and Automation Letters*, 7(2):2519–2526, 2022. 2

[12] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 3, 4

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[14] Partha Das, Sezer Karaoglu, and Theo Gevers. Intrinsic image decomposition using physics-based cues and cnns. *Computer Vision and Image Understanding*, 223:103538, 2022. 2

[15] Partha Das, Sezer Karaoglu, and Theo Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19799, 2022. 2, 4

[16] Joseph DeGol, Mani Golparvar-Fard, and Derek Hoiem. Geometry-informed material recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1554–1562, 2016. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[18] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8944–8952, 2018. 2, 4, 5, 6

[19] David Forsyth and Jason J Rock. Intrinsic image decomposition using paradigms. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7624–7637, 2021. 2

[20] Huan-ang Gao, Beiwen Tian, Pengfei Li, Xiaoxue Chen, Hao Zhao, Guyue Zhou, Yurong Chen, and Hongbin Zha. From semi-supervised to omni-supervised room layout estimation using point clouds. *arXiv preprint arXiv:2301.13865*, 2023. 2

[21] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7175–7183, 2019. 2

[22] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7154–7164, 2019. 3

[23] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009. 2

[24] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 112(2):133–149, 2015. 1

[25] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012. 3

[26] Hanzhe Hu, Yinbo Chen, Jiarui Xu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Learning implicit feature alignment function for semantic segmentation. *arXiv preprint arXiv:2206.08655*, 2022. 3

[27] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 187–203, 2018. 2

[28] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. *Advances in neural information processing systems*, 30, 2017. 2

[29] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017. 2

[30] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011. 2

[31] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 8

[32] Pengfei Li, Ruowen Zhao, Yongliang Shi, Hao Zhao, Jirui Yuan, Guyue Zhou, and Ya-Qin Zhang. Lode: Locally conditioned eikonal implicit scene completion from sparse lidar. *arXiv preprint arXiv:2302.14052*, 2023. 3

[33] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 2, 6

[34] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 371–387, 2018. 2, 6

[35] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. Openrooms: An end-to-end open framework for photorealistic indoor scene datasets. *arXiv preprint arXiv:2007.12868*, 2020. 5, 6

[36] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 4

[37] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3248–3257, 2020. 2

[38] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–217, 2018. 2

[39] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. Lime: Live intrinsic material estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6315–6324, 2018. 2

[40] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2019. 3

[41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 2, 5

[42] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 5

[43] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE international conference on computer vision*, pages 2992–2992, 2015. 2

[44] Takuya Narihira, Michael Maire, and Stella X Yu. Learning lightness from human judgement on relative reflectance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2965–2973, 2015. 2

[45] Thomas Nestmeyer and Peter V Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6789–6798, 2017. 2, 6

[46] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2, 3, 4

[47] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8843–8850, 2019. 2, 6

[48] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1, 3

[49] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008. 1

[50] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8598–8607, 2019. 6

[51] Li Shen, Ping Tan, and Stephen Lin. Intrinsic image decomposition with non-local texture cues. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008. 2

[52] Li Shen and Chuohao Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *CVPR 2011*, pages 697–704. IEEE, 2011. 2

[53] Li Shen, Chuohao Yeo, and Binh-Son Hua. Intrinsic image decomposition using a sparse representation of reflectance. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2904–2915, 2013. 2

[54] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 3

[55] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6926, 2019. 2

[56] Jiaxiang Tang, Xiaokang Chen, and Gang Zeng. Joint implicit image function for guided depth super-resolution. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4390–4399, 2021. 3, 4

[57] Beiwen Tian, Liyi Luo, Hao Zhao, and Guyue Zhou. Vibus: Data-efficient 3d scene parsing with viewpoint bottleneck and uncertainty-spectrum modeling. *ISPRS Journal of Photogrammetry and Remote Sensing*, 194:302–318, 2022. 2

[58] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 2

[59] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2314–2320, 2016. 2

[60] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv:1903.11816*, 2019. 3

[61] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercial rgbd cameras. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2508–2522, 2019. 2

[62] Qing Zhang, Jin Zhou, Lei Zhu, Wei Sun, Chunxia Xiao, and Wei-Shi Zheng. Unsupervised intrinsic image decomposition using internal self-similarity cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[63] Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, and Li Zhang. Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10–18, 2017. 2

[64] Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, and Li Zhang. Pointly-supervised scene parsing with uncertainty mixture. *Computer Vision and Image Understanding*, 200:103040, 2020. 2, 5, 6

[65] Haitian Zheng, Lu Fang, Mengqi Ji, Matti Strese, Yigitcan Özer, and Eckehard Steinbach. Deep learning for surface material classification using haptic and visual information. *IEEE Transactions on Multimedia*, 18(12):2407–2416, 2016. 2

[66] Yupeng Zheng, Chengliang Zhong, Pengfei Li, Huan-ang Gao, Yuhang Zheng, Bu Jin, Ling Wang, Hao Zhao, Guyue Zhou, Qichao Zhang, et al. Steps: Joint self-supervised nighttime image enhancement and depth estimation. *arXiv preprint arXiv:2302.01334*, 2023. 1

[67] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 1, 5

[68] Hao Zhou, Xiang Yu, and David W Jacobs. Glosh: Global-local spherical harmonics for intrinsic image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7820–7829, 2019. 2

[69] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *Proceedings of the IEEE international conference on computer vision*, pages 3469–3477, 2015. 2

[70] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2822–2831, 2022. 2, 5, 6

[71] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE international conference on computer vision*, pages 388–396, 2015. 2