Saliency Guided Depth Prediction from a Single Image

Yu Wang[1] and Lizhuang Ma[2]
[1,2]*Department of Computing, Shanghai Jiao Tong University*

Keywords: Single-image Depth Prediction, CNN.

Abstract: With the recent surge of deep neural networks, depth prediction from a single image has seen substantial progress. Deep regression networks are typically learned from large data without much constraints about the scene structure, thus often leading to uncertainties at discontinuous regions. In this paper, we propose a structure-aware depth prediction method based on two observations: depth is relatively smooth within the same objects, and it is usually easier to model relative depth than model the absolute depth from scratch. Our network first predicts an initial depth map and takes an object saliency map as input, which helps to teach the network to learn depth refinement. Specifically, a stable anchor depth is first estimated from the detected salient objects, and the learning objective is to penalize the difference in relative depth versus the estimated anchor. We show such saliency-guided relative depth constraint unveils helpful scene structures, leading to significant gains on the RGB-D saliency dataset NLPR and depth prediction dataset NYU V2. Furthermore, our method is appealing in that it is pluggable to any depth network and is trained end-to-end with no overhead of time during testing.

1 INTRODUCTION

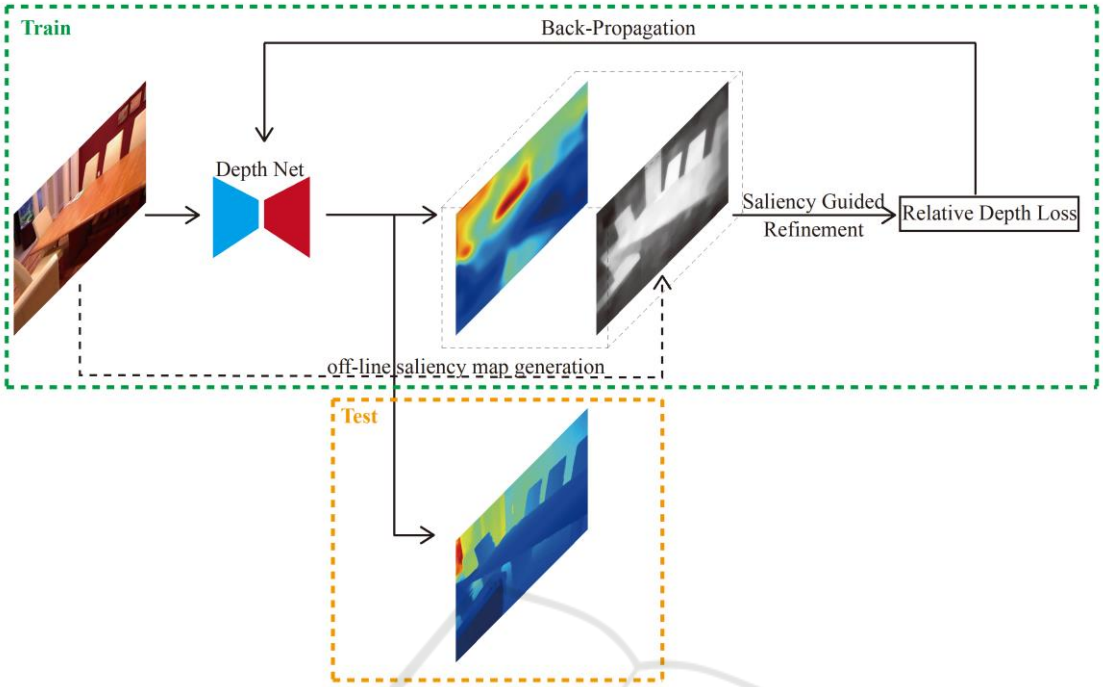
Depth prediction plays an essential role in understanding the 3D geometry of a scene. It has been shown that depth information can largely facilitate other vision tasks such as reconstruction (Silberman et al. 2012), recognition (Fox 2012) and semantic segmentation (Cheng et al. 2017). Stereo images (Kong and Black 2016) or image sequence (Suwajanakorn, Hernandez, and Seitz 2015) usually suffice for accurate depth prediction. While for single-view depth prediction, it is an ill-posed problem due to the lack of geometric information. Ambiguities or uncertainties often happen at those discontinuous regions between objects.

In this paper, we propose a structure-aware depth prediction method based on two observations: 1) depth is relatively smooth within the same objects when compared to within the full image; 2) and it is often easier to model relative depth than model the absolute depth value.

We incorporate such observations by learning to refine depth map with the guidance of object saliency (we use the saliency detector (Tong et al. 2015)). Generally, the saliency map of objects is a simple way to reveal the scene structure in terms of objects. Saliency map is also class-agnostic, thus can cover a broad range of objects each with spatially smooth depth values within its contour. As a result, we are able to first estimate an anchor depth value of the whole scene, from an initial depth map reweighted by object saliency. Since almost all object regions have small variance in depth values, such anchor depth estimation can act as a reliable reference. Then we take the anchor depth for depth refinements of entire image in a relative way. Previously (Chen et al. 2016) explored relative depth estimation, but both their depth ground truth and prediction are just ordinal, not real depth values. Here we design a relative depth loss for our network to learn the genuine relative depth of other pixels versus the estimated anchor depth, and penalize the deviation to correct relative depth.

Fig. 1 demonstrates our overall learning framework. During training, we propose two formulations of relative depth constraints to supervise the depth refinement process. At test time, such finetuned network is simply applied for depth prediction without any overhead. Obviously, our method is pluggable to any depth network and can be easily trained end-to-end. We show our saliency-guided depth model does learn some scene structures, leading to significant gains on the RGB-D saliency dataset NLPR and depth prediction dataset NYU V2.

Wang, Y. and Ma, L.
Saliency Guided Depth Prediction from a Single Image.
DOI: 10.5220/0008099101530159
In *Proceedings of the International Conference on Advances in Computer Technology, Information Science and Communications (CTISC 2019)*, pages 153-159
ISBN: 978-989-758-357-5



In summary, the contributions of our method are as follows:

- We propose the first single-image depth prediction method guided by object saliency, to the best of our knowledge.
- A novel relative depth loss is proposed to learn residuals versus salient anchor depth to improve prediction.
- Our method improves over state-of-the-art depth networks at no time cost while being pluggable to any type of depth prediction networks.

2 METHOD

2.1 Training Initialization

For each training image, we need to get the initial depth map prediction and a saliency map to improve the former. Our method is applicable to any depth network architecture, and we will show some popular ones in our experiments. The saliency map is acquired following the method in (Tong et al. 2015), which uses a bootstrap learning algorithm for salient object detection where both weak and strong models are exploited.

2.2 Anchor Depth Estimation

We already know that the depth is relatively smooth within object regions, which means their depth variance is pretty low compared to the variance in the full image. Hence we rely on object saliency detection to segment out all object regions with consistent depth values. We also find it reasonable to rely on the class-agnostic salient regions to obtain a weighted average of depth value as a stable reference in depth.

We consider the output depth map from the network as $\{d_i\}_{i=1}^{N}$, and the input saliency map as $\{p_i\}_{i=1}^{N}$, where p_i means the probability that i th pixel belongs to the salient object. After that, we define the saliency map as a weighting map and the normalized weights $\hat{p}$ as follows:

$$\hat{p}_i = \frac{p_i}{\sum_{i=1}^{N} p_i} \quad (1)$$

Figure 1: Overview of our saliency-guided depth prediction framework. One base depth netw
generates an initial depth map and we take an extra saliency map by (Tong et al. 2015) to refine de
first estimate a reliable anchor depth from object saliency-weighted depth values, and impose the
depth loss (versus anchor) as well as regular regression loss to finetune our network. This leads to i
structure-aware predictions, which can be obtained by simply forwarding the network at test time.

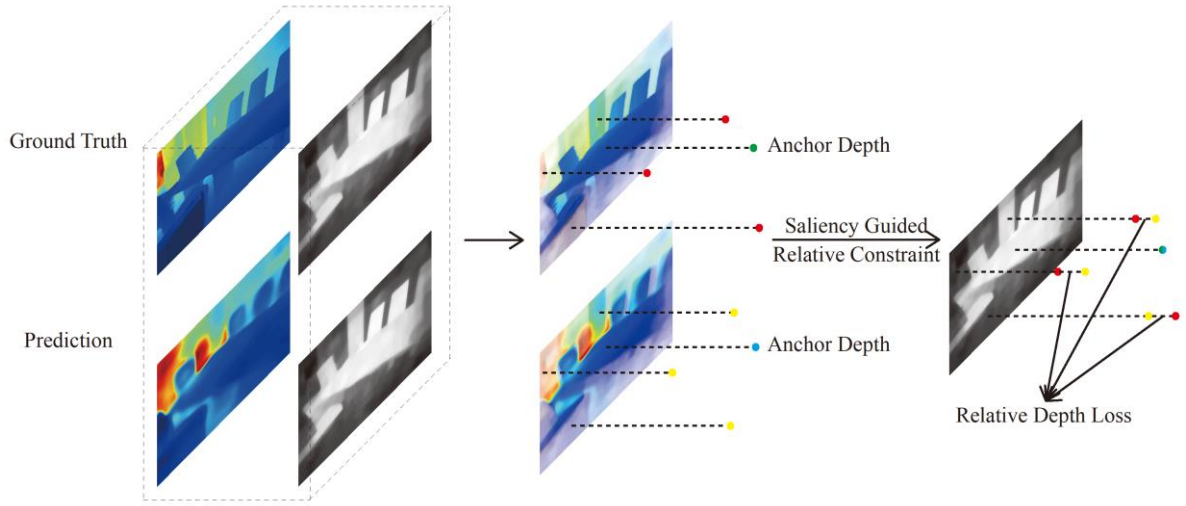


Figure 2: **Illustration of the Relative Depth Loss.** After estimating the anchor depth value, we normalize both the predicted and ground-truth depth maps by their absolute values within salient regions. We design two approaches to constrain relative depth relation corresponding to two kinds of depth ambiguities.

Then the depth conditional expectation of both prediction d and ground truth d^* on $\hat{p}$ can be formulated as:

$$E(d \mid \hat{p}) = \sum_{i=1}^{N} d_i * \hat{p}_i$$
$$E(d^* \mid \hat{p}) = \sum_{i=1}^{N} d_i^* * \hat{p}_i \quad (2)$$

Through these calculations, we obtain the baseline depth values of the prediction and the ground truth, which are relatively more reliable than the global mean values: $Var(d \mid p^*) << Var(d)$.

2.3 Relative Depth Loss

To define the relative depth loss, we first normalize the prediction and the ground truth with the anchor depth value as shown in Figure 2. Here we design two kinds of relative depth constraints, corresponding to two common types of monocular ambiguities. By means of these, the accurate prediction can be propagated from the reliable salient regions.

2.3.1 Absolute Difference Formulation

One representation of monocular ambiguity is that the absolute difference of depth values between different regions of a single image can be hardly confirmed. To deal with this kind of ambiguity, we define the depth absolute relation both over the prediction and the ground truth as follows:

$$r_{abs} = d - E(d \mid \hat{p})$$
$$r_{abs}^* = d - E(d^* \mid \hat{p}) \quad (3)$$

The depth absolute relation measures depth differences as if both the prediction and the ground truth values are shift along the optical line of sight regardless of the bias between two expectations calculated in Equation (2).

2.3.2 Relative Ratio Formulation

Another representation of monocular ambiguity is that the depth ratio between different regions of a single image can be hardly judged. Likewise we define the depth ratio relation like this:

$$r_{ratio} = d / E(d \mid \hat{p})$$
$$r_{ratio}^* = d / E(d^* \mid \hat{p}) \quad (4)$$

By the constraint towards depth ratio relations, we can exact the depth scale between the prediction and the ground truth as if the two depth maps are projected and aligned with their salient regions.

2.4 Full Objective

Generally speaking, a standard loss function for non-parametric regression problems is L norm loss (usually MSE) or even some advantage regression loss(*i.e.* BerHu from (Laina et al. 2016)) between the prediction d and ground truth depth value d^*. We record this loss as L_d.

Besides the standard depth loss L_d above, we choose L1 loss as the criterion over two kinds of depth relations. That is:

$$L_{abs} = E(|r_{abs} - r^*_{abs}|)$$
$$L_{ratio} = E(|r_{ratio} - r^*_{ratio}|) \quad (5)$$

And our final loss function can be formulated as follows:

$$L = L_d + \lambda_{abs} * L_{abs} + \lambda_{ratio} * L_{ratio} \quad (6)$$

Here λ_{abs} and λ_{ratio} act as the loss weights.

3 EXPERIMENTS

3.1 Datasets

To evaluate the effectiveness of our method, we perform experiments over two datasets: the RGB-D saliency dataset NLPR (Peng et al. 2014) and indoor depth prediction benchmark dataset NYU depth v2 (Silberman et al. 2012).

3.1.1 NLPR Dataset

The NLPR dataset consists of 1000 RGB-D pairs collected with Kinect, whose salient object is already labeled out. The aspect ratios of data in this dataset are not uniformed, with 698 horizontal and 302 vertical. To make a fair comparison, we randomly choose a test set with 245 horizontal and 108 vertical, which shares almost the same proportion between different aspect ratios.

3.1.2 NYU Dataset

The NYU depth dataset is one of the largest RGB-D datasets for indoor scene reconstruction. The labeled part contains 1449 RGB-D pairs with dense depth maps. We perform this part of experiments on the labeled part of this dataset and use the common test subset of 654 images split by (Eigen, Puhrsch, and Fergus 2014).

3.2 Metrics

We evaluate each method using several errors and accuracy settings from previous depth prediction works:

Abs Rel $\quad \frac{1}{N}\sum_i \frac{|d_i - d_i^*|}{d_i^*}$

$\log_{10}$ $\quad \frac{1}{N}\sum_i |\log_{10} d - \log_{10} d^*|$

RMSE $\quad \sqrt{\frac{1}{N}\sum_i |d_i - d_i^*|^2}$

Accuracy $\quad$ % of d_i s.t. $\max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) < \delta$

3.3 Results

3.3.1 NLPR Dataset

Quantitative results for NLPR dataset are provided in Table 1. It is illustrated that our method benefits each baseline method both on accuracies and errors. From the qualitative results shown in Figure 3, we find that through our method the structure of the salient region could be preserved compared to the baseline method. And the performance of our method would be limited by the accuracy of the saliency map. The structure of salient object in the vertical case are not so clear compared to the horizontal case, as the quality of these two saliency maps vary obviously. The effects by the accuracy of saliency maps are further evaluated in the Ablation Study subsection.

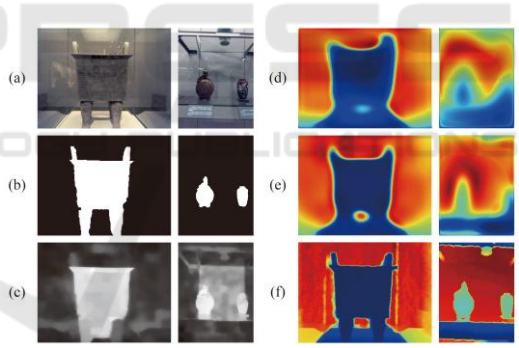


Figure 3: **Qualitative results of the NLPR dataset.** (a): input images. (b): ground truth saliency labels. (c): saliency maps generated off-line following the method in (Tong et al. 2015). (d): output depth maps of the best baseline method (Laina et al. 2016)*. (e): output depth maps of our proposed method.

3.3.2 NYU Dataset

The quantitative results in Table 2 illustrate that our method help each baseline method to gain an overall boost in performance. Although the ability of pure-learning baseline methods is relatively limited with small amount of training data, our proposed method cascading on baselines would be comparable with

other state-of-the-art methods. Note that the NYU Depth dataset actually does not contain salient object label, but the saliency maps we obtained are still meaningful.

Table 1: **Quantitative results of the NLPR dataset**. $\delta_i < 1.25^i$. * means ResNet-50 based model.

	higher is better			lower is better		
Method	δ_1	δ_2	δ_3	Abs Rel	$\log_{10}$	RMSE
(Eigen, Puhrsch, and Fergus 2014)	0.176	0.387	0.622	0.692	0.255	1.410
(Laina et al. 2016)	0.281	0.537	0.701	0.479	0.225	1.321
(Laina et al. 2016)*	**0.466**	**0.742**	**0.886**	**0.344**	**0.141**	**0.900**
Cascade on (Eigen, Puhrsch, and Fergus 2014)	0.272	0.527	0.693	0.474	0.233	1.350
Cascade on (Laina et al. 2016)	0.386	0.655	0.818	0.443	0.173	1.093
Cascade on (Laina et al. 2016)*	**0.504**	**0.764**	**0.897**	**0.321**	**0.132**	**0.868**

Table 2: **Quantitative results on the NYU Depth dataset**. $\delta_i < 1.25^i$. * means ResNet-50 based model.

	higher is better			lower is better		
Method	δ_1	δ_2	δ_3	Abs Rel	$\log_{10}$	RMSE
(Saxena, Sun, and Ng 2009)	0.447	0.745	0.897	0.349	-	1.214
(Karsch, Liu, and Kang 2012)	-	-	-	0.350	0.131	1.200
(Liu, Salzmann, and He 2014)	-	-	-	0.335	0.127	1.06
(Li et al. 2015)	0.621	0.886	0.968	0.232	0.094	0.821
(Wang et al. 2015)	0.605	0.906	0.970	0.220	-	0.824
(Liu et al. 2016)	0.650	0.906	0.976	0.213	0.087	0.759
(Eigen, Puhrsch, and Fergus 2014)	0.498	0.798	0.929	0.312	0.122	0.946
(Laina et al. 2016)	0.587	0.859	0.956	0.255	0.102	0.805
(Laina et al. 2016)*	**0.610**	**0.875**	**0.961**	**0.252**	**0.097**	**0.760**
Cascade on (Eigen, Puhrsch, and Fergus 2014)	0.539	0.826	0.938	0.291	0.113	0.888
Cascade on (Laina et al. 2016)	0.618	0.875	0.960	0.246	0.096	0.760
Cascade on (Laina et al. 2016)*	**0.651**	**0.897**	**0.970**	**0.229**	**0.090**	**0.696**

Table 3: **Ablation studies on the NLPR dataset**. $\delta_i < 1.25^i$. Backbones are uniformed to ResNet-50. 'A' means the constraint of depth absolute relation. 'R' means the constraint of depth ratio relation. 'F' means full-image reweighting. '0.XF' means foreground reweighting with the threshold as 0.X. 'GT' means this experiment use the ground truth saliency maps to reweight depth maps. Other implementation details are the same with the experiments mentioned above.

	higher is better			lower is better			Saliency
Method	δ_1	δ_2	δ_3	Abs Rel	$\log_{10}$	RMSE	Cos Similarity
Res50	0.466	0.742	0.886	0.344	0.141	0.900	0.648
Res50-A	0.503	0.758	0.893	0.332	0.134	0.872	0.648
Res50-R	**0.505**	0.760	0.893	0.332	0.134	0.870	0.648
Res50-A-R(proposed)	0.504	**0.764**	**0.897**	**0.321**	**0.132**	**0.868**	0.648
Res50-A-R-F	0.513	0.766	0.896	0.327	0.132	0.874	0.333
Res50-A-R-0.3F	0.508	0.761	0.894	0.335	0.132	0.871	0.352
Res50-A-R-0.4F	0.512	0.767	0.895	0.341	0.132	0.879	0.388
Res50-A-R-0.5F	0.512	0.767	0.894	0.338	0.132	0.874	0.411
Res50-A-R-GT	**0.580**	**0.784**	**0.898**	0.328	0.138	**0.857**	1.000

3.4 Ablation Study

We conduct a series of ablation studies to analyze the details of our method on both datasets. Quantitative results are shown in Table 3.

3.4.1 Mode of Anchor Depth Estimation

In our method we use the saliency map to reweight both the prediction and ground truth to confirm the anchor depth. To confirm the effects of this reweighting, we design two other reweighting ways: full-image reweighting and foreground reweighting. Full-image reweighting makes the weight of every pixel p_i equals to 1. Foreground reweighting makes the weights equals to 1 on pixels whose depth value ranks front within the threshold percentage and leaves the rest to 0.

The results on both datasets indicate that different reweighting ways have some effects but our reweighting based on salient object detection is more effective than alternative ways.

3.4.2 Constraints of Depth Relations

The two kinds of constraints towards depth relations are designed for different representations of monocular ambiguity. Thus we evaluate the gain of each kind independently.

The results on both datasets show that two kinds of constraints have their own positive effects and would gain advance when performed together.

3.4.3 Accuracy of Saliency Map

The saliency maps used in our method are generated off-line following (Tong et al. 2015). We explore the effects of saliency accuracy for the saliency ground truth is contained in the NLPR dataset. According to Equation (1) in section 2.2, the saliency map only have proportional meaning so we evaluate saliency accuracy with cosine similarity.

Results in Table 3 show that the accuracy of saliency would slightly contribute to lower RMSE and higher δ_3, but the overall effect is difficult to evaluate. One possible reason is that the definition of saliency is subjective so that the confidence of the accuracy remains some doubt. However, this does not fade the effectiveness of our method.

4 CONCLUSIONS

This paper proposes a structure-aware deep model for depth prediction from single image. Our network predicts an initial depth map and uses an extra object saliency map to learn depth refinement. Concretely, we estimate a stable anchor depth value from the detected salient depth regions, and learn to penalize the difference in relative depth versus the estimated anchor. We show such relative depth constraint significantly improves depth prediction accuracy, and learns some helpful scene structures. Furthermore, our method is pluggable to any depth network and can be trained end-to-end without inference overhead.

REFERENCES

Chen, W.; Fu, Z.; Yang, D.; and Deng, J. 2016. Single-image depth perception in the wild. In Lee, D. D.; Sugiyama, M.;Luxburg, U. V.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc. 730 – 738

Eigen, D., and Fergus, R. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2650 – 2658.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 2366 – 2374.

Fox, D. 2012. Rgb-(d) scene labeling: Features and algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2759 – 2766.

Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2016. Unsupervised monocular depth estimation with left-right consistency. 6602 – 6611.

Karsch, K.; Liu, C.; and Kang, S. B. 2012. Depth extraction from video using non-parametric sampling. *In Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV'12, 775 – 788. Berlin, Heidelberg: Springer-Verlag.

Kuznietsov, Y.; Stueckler, J.; and Leibe, B. 2017. Semisupervised deep learning for monocular depth map prediction. In *cvpr*.

Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; and Navab, N. 2016. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV),*

2016 Fourth International Conference on, 239 – 248. IEEE.

Li, B.; Shen, C.; Dai, Y.; van den Hengel, A.; and He, M. 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In The *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, M.; Salzmann, M.; and He, X. 2014. Discretecontinuous depth estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 716–723.

Peng, H.; Li, B.; Xiong, W.; Hu, W.; and Ji, R. 2014. Rgbd salient object detection: a benchmark and algorithms. In *European Conference on Computer Vision (ECCV)*, 92–109.

Saxena, A.; Sun, M.; and Ng, A. Y. 2009. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell*. 31(5):824–840.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. 7576(1):746–760.

Suwajanakorn, S.; Hernandez, C.; and Seitz, S. M. 2015. Depth from focus with your mobile phone. In *CVPR*, 3497–3506. IEEE Computer Society.

Tong, N.; Lu, H.; Ruan, X.; and Yang, M. 2015. Salient object detection via bootstrap learning. In 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1884–1892.

Wang, P.; Shen, X.; Lin, Z.; Cohen, S.; Price, B. L.; and Yuille, A. L. 2015. Towards unified depth and semantic prediction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2800–2809.