Monitored Distillation for Positive Congruent Depth Completion

Tian Yu Liu[1⋆], Parth Agrawal[1⋆], Allison Chen[1⋆], Byung-Woo Hong[2], and Alex Wong[3]

[1] UCLA Vision Lab, Los Angeles, CA 90095, USA
{tianyu139,parthagrawal24,allisonchen2}@ucla.edu
[2] Chung-Ang University, Heukseok-Dong, Dongjak-Gu, Seoul, 06973, Korea
hong@cau.ac.kr
[3] Yale University, New Haven, CT 06511, USA
alex.wong@yale.edu

Abstract. We propose a method to infer a dense depth map from a single image, its calibration, and the associated sparse point cloud. In order to leverage existing models (teachers) that produce putative depth maps, we propose an adaptive knowledge distillation approach that yields a positive congruent training process, wherein a student model avoids learning the error modes of the teachers. In the absence of ground truth for model selection and training, our method, termed *Monitored Distillation*, allows a student to exploit a blind ensemble of teachers by selectively learning from predictions that best minimize the reconstruction error for a given image. Monitored Distillation yields a distilled depth map and a confidence map, or "monitor", for how well a prediction from a particular teacher fits the observed image. The monitor adaptively weights the distilled depth where if all of the teachers exhibit high residuals, the standard unsupervised image reconstruction loss takes over as the supervisory signal. On indoor scenes (VOID), we outperform blind ensembling baselines by 17.53% and unsupervised methods by 24.25%; we boast a 79% model size reduction while maintaining comparable performance to the best supervised method. For outdoors (KITTI), we tie for 5th overall on the benchmark despite not using ground truth. Code available at: https://github.com/alexklwong/mondi-python.

Keywords: depth completion, blind ensemble, knowledge distillation

1 Introduction

Interaction with physical space requires a representation of the 3-dimensional (3D) geometry of the surrounding environment. Most mobile platforms include at least one camera and some means of estimating range at a sparse set of points i.e. a point cloud. These could be from a dedicated range sensor such as a LiDAR or radar, or by processing the images using a visual odometry module.

⋆ denotes equal contribution.

Depth completion consists of inferring a dense depth map, with a range value corresponding to every pixel, from an image and a sparse point cloud. Inherently, depth completion is an ill-posed inverse problem, so priors need to be imposed in the form of generic regularization or learned inductive biases.

Natural scenes exhibit regularities that can be captured by a trained model, for instance a deep neural network (DNN), using a dataset of images and corresponding sparse depths. While we wish to avoid any form of manual or ground truth supervision, we also strive to exploit the availability of differing types of pretrained models, whether from synthetic data or other supervised or unsupervised methods. We refer to these pretrained models as "teachers," each providing a hypothesis of depth map for a given image and sparse point cloud. This leads to a blind ensemble setting where ground truth is not available (e.g. transferring models trained on a specific task to new datasets with no ground truth) for the explicit evaluation of pretrained models i.e. model selection. The key question, then, is how to make use of a heterogeneous collection of teachers, along with other variational principles such as minimization of the photometric reprojection error and generic regularizers such as structural similarity.

In general, different teachers will behave differently not only across images, but even across regions within a given image. The incongruency of different models trained on the same tasks has been observed in the context of classification model versioning [61]. Particularly, the same architecture trained with the same data, but starting from different initial conditions can yield models that *differ on a significant portion of the samples* while achieving the same average error rate. Thus, a naive ensembling of a handful of teachers yields the union of the failure modes, only modestly mitigated by the averaging.

Instead, we propose Monitored Distillation for selecting which teacher to emulate in each image at each pixel. The selection is guided by a "monitor", based on the residual between the observations (e.g. image, sparse point cloud) and their reconstructions generated by each teacher. This yields a spatially-varying confidence map that weights the contribution of the selected teachers as well as the structural and photometric reprojection errors i.e. unsupervised losses, customary in structure-from-motion. In doing so, our method is robust even when poor performing teachers are introduced into the ensemble – discarding their hypotheses in favor of the ones that better reconstruct the scene. In the extreme case where every teacher produces erroneous outputs, our method would still learn a valid depth estimate because of our unsupervised fall-back loss.

Our contributions are as follows: **(i)** We propose an adaptive method to combine the predictions of a blind ensemble of teachers based on their compatibility with the observed data; to the best of our knowledge, we are the first to propose knowledge distillation from a blind ensemble for depth completion. **(ii)** The adaptive mechanism yields a spatially varying confidence map or "monitor" that modulates the contributions of each teacher based on their residuals, leading to a training method that is positive congruent. **(iii)** Even when all members of the ensemble fail, our model automatically reverts to the unsupervised learning criteria and generic regularization, allowing us to avoid distilling erroneous

knowledge from teachers. **(iv)** Our method outperforms distillation and unsupervised methods by 17.53% and 24.25% respectively on indoors scenes; we are comparable to top supervised methods with a 79% model size reduction. On the KITTI benchmark, we tie for 5th overall despite not using ground truth.

2 Related Works

Depth completion is a form of imputation and thus requires regularization, which may come from generic assumptions or learned from data. The question is: How to best combine different sources of regularization, adaptively [17,18,57], in a way that leverages their strengths, while addressing their weaknesses?

Supervised depth completion is trained by minimizing a loss with respect to ground truth. Early methods posed the task as learning morphological operators [10] and compressive sensing [8]. Recent works focus on network operations [12,22] and design [5,35,49,62] to effectively deal with the sparse inputs. For example, [29] used a cascade hourglass network, [24,62] used separate image and depth encoders and fused their representations, and [22] proposed an upsampling layer and joint concatenation and convolution. Whereas, [11,12,43,44] learned uncertainty of estimates, [50] leveraged confidence maps to fuse predictions from different modalities, and [42,60,63] used surface normals for guidance. [6,39] use convolutional spatial propagation networks, [21] used separate image and depth networks and fused them with spatial propagation. While supervised methods currently hold the top ranks on benchmark datasets i.e. KITTI [49] and VOID [56], they inevitably require ground truth for supervision, which is typically unavailable. Furthermore, these architectures are often complex and require many parameters (e.g. 132M for [21], 53.4M for [42], and 25.8M for [39]), making them computationally prohibitive to train and impractical to deploy [37].

Unsupervised depth completion assumes that additional data (stereo or monocular videos) is available during training. Both stereo [46,62] and monocular [35,54,55,56] paradigms focus largely on designing losses that minimize (i) the photometric error between the input image and its reconstructions from other views, and (ii) the difference between the prediction and sparse depth input (sparse depth reconstruction). Architecture-wise, [58] proposed a calibrated backprojection network. However, all of these methods rely on *generic* regularization i.e. local smoothness that is not informed by the data. Attempts to leverage learned priors mainly focused on synthetic data. [34] applied image translation to obtain ground truth in the real domain; whereas [56,62] used synthetic data to learn a prior on the shapes populating a scene. We also employ an unsupervised loss, but unlike them, we distill regularities from a blind ensemble of pretrained models that can be trained on synthetic or real data, supervised or unsupervised.

Knowledge Distillation uses a simpler student model to approximate the function learned by a larger, more complex teacher model by training it to learn the soft target distribution [16]. There exists many works on knowledge distillation, including image classification [33,45,59], object detection [2,3,4], semantic segmentation [31,38,40], depth estimation [20,32,52], and more recently, depth

completion [23]. [9,31,32] utilize pairwise and holistic distillation to capture structural relationships and [38] distills latent representations to guide learning. [52] leverages knowledge distillation for monocular depth estimation on mobile devices, and [41] uses cyclic inconsistency and knowledge distillation for unsupervised depth estimation, where the student network is a sub-network of the teacher. In depth completion, [23] uses knowledge distillation for joint training of both teacher and student models. Unlike ours, this method uses ground truth.

Ensemble learning addresses the limitations of a single teacher by distilling information from multiple teachers [14]. If done effectively, the student will learn to extract the most relevant information from each teacher. This has been explored in classification [26,28,51], but fewer works utilize it in dense prediction tasks. [1] uses it for domain adaptation in semantic segmentation and [15] in selecting lidar points for depth completion. We further assume the blind ensemble setting [48] where we lack ground truth for evaluation of the ensemble.

Positive congruent training [61] observed sample-wise inconsistencies in classification versioning, where new models wrongly predict for samples that were previously classified correctly by an older, reference model on the same task and dataset. To address this, they propose to emulate the reference model (teacher) only when its predictions are correct; otherwise, they minimize a loss with respect to ground truth – yielding reduced error rates and inconsistencies. Monitored distillation is inspired by positive-congruency, but unlike [61], we do not require ground truth and are applicable towards geometric tasks.

3 Method Formulation

We wish to recover the 3D scene from a calibrated RGB image $I : \Omega \subset \mathbb{R}^2 \mapsto \mathbb{R}^3_+$ and its associated sparse point cloud projected onto the image plane $z : \Omega_z \subset \Omega \mapsto \mathbb{R}_+$. To do so, we propose learning a function f_θ that takes as input I, z, and camera intrinsics K and outputs a dense depth map $\hat{d} := f_\theta(I, z, K) \in \mathbb{R}_+^{H \times W}$.

We assume that for each synchronized pair of image and sparse depth map (I_t, z_t) captured at a viewpoint t, we have access to a set of spatially and/or temporally adjacent alternate views T and the corresponding set of images I_T. Additionally, we assume access to a set of M models or "teachers" $\{h_i\}_{i=1}^M$ (e.g. publicly available pretrained models). Fig. 2 shows that each teacher has unique failure modes. As we operate in the blind ensemble setting, we lack ground truth to evaluate teacher performance for model selection. To address this, we propose Monitored Distillation, an adaptive knowledge distillation framework for ensemble learning that results in positive congruent training: We only learn from a teacher if its predictions are compatible with the observed scene.

To this end, we leverage geometric constraints between I_t and $I_\tau \in I_T$ and validate the correctness of predictions $d_i := h_i(I_t, z_t)$ produced by each teacher through averaging their photometric reprojection residuals from different views I_T and weighting them based on deviations from z. From the error, we derive a confidence map that determines the compatibility of each teacher to the observed image I_t. We then construct distilled depth $\bar{d} \in \mathbb{R}_+^{H \times W}$ via pixel-wise

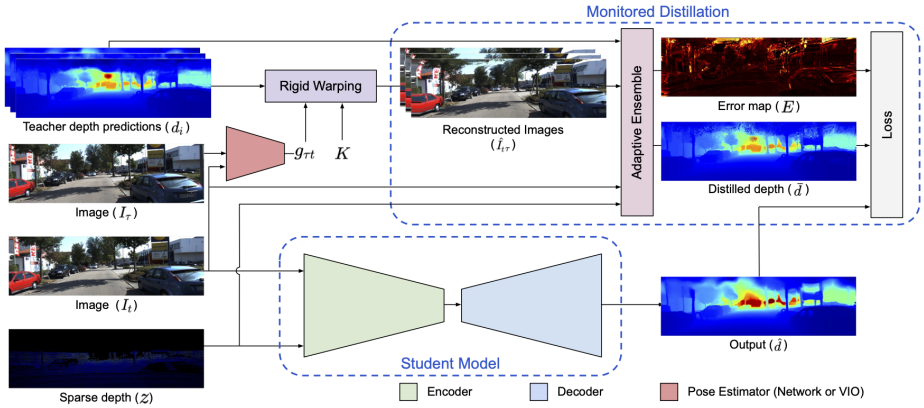


Fig. 1: **Monitored Distillation.** Our method measures the reconstruction residual of predictions from each teacher and constructs the distilled depth $\bar{d}$ based on a pixel-wise selection of predictions that best minimize the reconstruction error E. We derive a monitor function Q from E, which adaptively balances the trade-offs between the distilling from the ensemble and the unsupervised losses.

selection from the ensemble that yields the highest confidence. The resulting spatially varying confidence map acts as a "monitor" to balance the trade-off between "trusting" the ensemble and falling back onto unsupervised geometric consistency as a supervisory signal (i.e. when all teachers yield high residuals).

Monitored Distillation. Given $M \in \mathbb{Z}^+$ teachers and their predicted depth maps d_i, for $i \in \{1, \cdots, M\}$, we construct a distilled depth map $\bar{d}$ by adaptively selecting predictions from the teacher ensemble that best minimize reconstruction error of the observed point cloud and image. To this end, we reconstruct the observed image I_t via reprojection from an adjacent view $I_\tau, \tau \in T$:

$$\hat{I}_{t\tau}(x, d) = I_\tau(\pi g_{\tau t} K^{-1} \bar{x} d(x)), \tag{1}$$

where d denotes the depth values for $x \in \Omega$, $\bar{x} = [x^\top \ 1]^\top$ is the homogeneous coordinate of x, $g_{\tau t} \in SE(3)$ is the relative pose of the camera from view t to τ, K is the camera intrinsics, and π is the perspective projection. In practice, $g_{\tau t}$ can be derived from camera baseline if I_t and I_τ are stereo pairs, directly estimated by a visual inertial odometry (VIO) system, or learned from a pose network if the views are taken from a video.

For each teacher h_i, we measure the photometric reprojection error P_i via the mean SSIM [53] between I_t and each reconstruction $\hat{I}_{t\tau}(x, d_i), \tau \in T$:

$$P_i(x) = \frac{1}{|T|} \sum_{\tau \in T} \left(1 - \phi(\hat{I}_{t\tau}(x, d_i), I_t(x))\right) \tag{2}$$

For ease of notation, we denote SSIM as $\phi(\cdot)$.

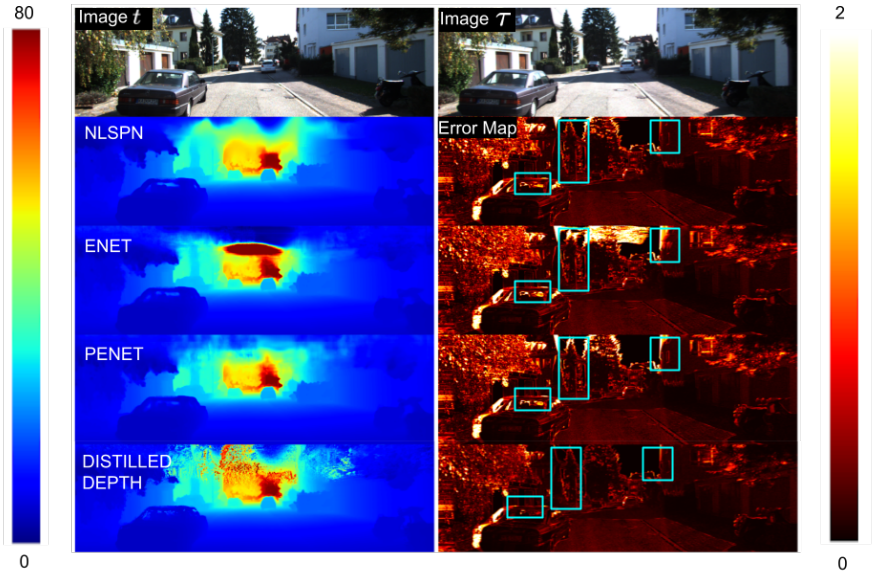


Fig. 2: **Error Modes of Teacher Models on KITTI**. Row 1 shows the input image I_t (left) and an image taken from another view I_τ (right). Rows 2-4 shows the predicted depth maps (left) and error maps (right) for each teacher, where each has different error modes. Row 5 shows the distilled depth that our method adaptively constructs from the teacher models. The error for each region in the distilled depth lower bounds the reprojection error of the individual teachers.

As photometric reprojection alone does not afford scale, we additionally measure the local deviation of teacher predictions from the observed sparse point cloud within a $k \times k$ neighborhood of x, denoted by $\mathcal{N}(x)$:

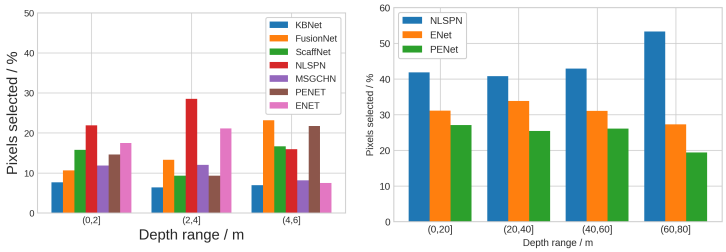
$$Z_i = \frac{1}{k^2|z_t|} \sum_{x \in \Omega} \sum_{y \in \mathcal{N}(x)} \mathbb{1}_{z_t(x)} \cdot |d_i(y) - z_t(x)| \tag{3}$$

When used as a weight, $\beta_i := 1 - \exp(-\alpha Z_i)$ serves to resolve the scale ambiguity between different teachers, where α is a temperature parameter. We can then define E_i, the weighted reconstruction residual from the i-th teacher, as:

$$E_i(x) = \beta_i P_i(x). \tag{4}$$

To construct the distilled depth, we selectively choose the depth prediction for each pixel $x \in \Omega$ that minimizes the overall residual error E_i across all teachers:

$$\bar{d}(x) = \sum_{i=1}^{M} \mathbb{1}_i(x) d_i(x), \tag{5}$$



(a) VOID Dataset (indoor) (b) KITTI Dataset (outdoor)

Fig. 3: **Teacher Selection Distribution.** The plots show the proportion of pixels selected from each teacher model. Note: error modes vary across different depth ranges as different teachers dominate the selection at different distances.

where $\mathbb{1}_i$ is a binary weight map of the i-th teacher is given by

$$\mathbb{1}_i(x) = \begin{cases} 1 & E_i(x) < E_j(x) \;\forall\; j \neq i \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

In other words, $\mathbb{1}_i(x) = 1$ when $d_i(x)$ yields the lowest photometric residual. Fig. 1 shows an overview of our method, where teacher predictions are ensembled into distilled depth for supervision. Fig. 3 shows the distribution of teachers chosen for constructing the distilled depth. As observed, different teachers perform well in different regions across different depth ranges. Our method selects points from each teacher with the lowest error (i.e. highest confidence) to yield an adaptive ensemble (see Fig. 2).

Despite being trained on ground truth, each teacher can only *approximate* the true distribution of depths in the scene. While selectively ensembling based on the reprojection residual will address *some* error modes of the teachers, it is still possible for all teachers to yield high reconstruction residuals. Hence, we do not trust the ensemble fully, and instead further adaptively weight the ensemble supervision with a monitor Q based on the error of the distilled depth $\bar{d}$. As we have already constructed the error maps E_i for each teacher, we can similarly aggregate the error for each pixel $E(x) = \min_i E_i(x)$ for $x \in \Omega$. The final monitor $Q \in [0,1]^{H \times W}$ is a spatially adaptive per-pixel confidence map:

$$Q(x) = \exp(-\lambda E(x)), \tag{7}$$

where λ is a temperature parameter (see Supp. Mat.). Q naturally assigns higher confidence to points in the distilled depth that are compatible with the observed image I_t as measured by reconstruction error, and is used to weight the supervision signal. Our monitored knowledge distillation objective reads:

$$\ell_{md} = \frac{1}{|\Omega|} \sum_{x \in \Omega} Q(x) \cdot |\hat{d}(x) - \bar{d}(x)|. \tag{8}$$

Typically, a student learns the error modes of its teacher. But by distilling from the adaptive ensemble of teachers that is positive congruent, our student model learns not to make the same mistake as any individual teacher. We refer to this process as *Monitored Distillation*, in which our monitor function Q gives higher weight to the teachers in regions of lower reconstruction error. For regions where all of the teachers within the ensemble yield high residuals, we default to unsupervised loss to avoid learning the common error modes of any teacher.

Unsupervised Objective. For regions with high reconstruction error, the monitoring function Q allows us to fall back onto standard unsupervised photometric reprojection error, i.e. color and structural consistencies, as the training signal:

$$\ell_{co} = \frac{1}{|\Omega|}\frac{1}{|T|}\sum_{x\in\Omega}\sum_{\tau\in T}(1-Q(x))\big(|\hat{I}_{t\tau}(x,\hat{d}) - I_t(x)|\big) \tag{9}$$

$$\ell_{st} = \frac{1}{|\Omega|}\frac{1}{|T|}\sum_{x\in\Omega}\sum_{\tau\in T}(1-Q(x))\big(1 - \phi(\hat{I}_{t\tau}(x,\hat{d}), I_t(x))\big) \tag{10}$$

We weight the relative contributions of these losses with the complement of our adaptive monitor function $(1-Q)$. As a result, our framework naturally allows us to search for the correct correspondences (and in turn better depth estimation) in regions where the ensemble failed. In other words, regions which the monitor deems as high confidence are more heavily influenced by ℓ_{md} as supervision, while lower confidence regions will minimize unsupervised losses instead.

Because the ensemble is informed by large amounts of data, their predictions have regularities of our physical world, e.g. roads are flat and surfaces are locally connected, "baked into" them. This presents an advantage: The student will learn priors, often too complex to be modeled by generic assumptions, from the ensemble. However, these priors may backfire when all the teachers yield high residuals. Luckily, Q naturally limits the influence of the ensemble in such cases, but this in turn reduces the amount of regularization that is needed for ill-posed problems like 3D reconstruction. Hence, for these cases, we default to generic assumptions i.e. a local smoothness regularizer:

$$\ell_{sm} = \frac{1}{|\Omega|}\sum_{x\in\Omega}(1-Q(x))\big(\lambda_X(x)|\partial_X\hat{d}(x)| + \lambda_Y(x)|\partial_Y\hat{d}(x)|\big) \tag{11}$$

where ∂_X, ∂_Y are gradients along the x and y directions, weighted by $\lambda_X := e^{-|\partial_X I_t(x)|}$ and $\lambda_Y := e^{-|\partial_Y I_t(x)|}$ respectively.

Thus, we have the following overall loss function

$$\mathcal{L} = w_{md}\ell_{md} + w_{ph}\ell_{ph} + w_{st}\ell_{st} + w_{sm}\ell_{sm} \tag{12}$$

where $w_{(\cdot)}$ denotes the respective weights for each loss term (see Supp. Mat.).

Student Model Architecture. Through monitored distillation from an ensemble of teachers, a simpler student model can be trained on the output distribution

Table 1: **Blind Ensemble Distillation.** We compare Monitored Distillation against naive ensembling methods for training a student model.

Ensemble Type	Distillation Method	MAE	RMSE	iMAE	iRMSE
None	Unsupervised Loss Only	55.67	117.21	28.68	58.31
Supervised	Mean w/o Unsupervised Loss	34.27	91.72	17.63	41.39
	Mean	34.04	89.19	17.30	40.43
	Median	34.64	89.80	17.46	39.77
	Random	35.18	92.30	18.41	42.95
	Ours w/o β	32.86	85.53	16.44	39.14
	Ours	**30.88**	**87.48**	**15.31**	**38.33**
Unsupervised	Mean w/o Unsupervised Loss	44.73	96.56	24.08	49.55
	Mean	41.96	94.47	23.80	50.37
	Median	43.86	99.46	23.62	50.85
	Random	39.38	92.14	20.62	46.04
	Ours w/o β	38.78	90.72	20.53	45.91
	Ours	**36.42**	**87.78**	**19.18**	**43.83**
Heterogeneous	Mean w/o Unsupervised Loss	44.53	100.59	23.33	48.36
	Mean	35.79	84.78	18.65	42.90
	Median	33.89	85.25	17.31	40.40
	Random	43.64	94.38	24.74	50.27
	Ours w/o β	32.09	80.20	16.15	38.86
	Ours	**29.67**	**79.78**	**14.84**	**37.88**

of more complex teacher models to achieve comparable performance. Therefore, we compress KBNet [58] by replacing the final two layers in the encoder with depth-wise separable convolutions [19] to yield a 23.2% reduction in the number of model parameters. Compared to the best supervised teacher models that require 25.84M (NLSPN [39]), 131.7M (ENet [21]), 132M (PENet [21]), and 6.9M (the original KBNet) parameters, our student model only requires 5.3M.

4 Experiments

We evaluate our method on public benchmarks – VOID [56] for indoor and outdoor scenes and KITTI [49] for outdoor driving settings. We describe evaluation metrics, implementation details, hyper-parameters and learning schedule in the Supp. Mat. [30]. All experiments are performed under the blind ensemble setting where we do not have ground truth for model selection nor training.

VOID dataset [56] contains synchronized 640×480 RGB images and sparse depth maps of indoor (laboratories, classrooms) and outdoor (gardens) scenes. The associated sparse depth maps contain $\approx$1500 sparse depth points with a density of $\approx$0.5%. They are obtained by a set of features tracked by XIVO [13], a VIO system. The dense ground-truth depth maps are acquired by active stereo. As opposed to static scenes in KITTI, the VOID dataset contains 56 sequences with challenging motion. Of the 56 sequences, 48 sequences ($\approx$45,000 frames) are designated for training and 8 for testing (800 frames). We follow the evaluation protocol of [56] and cap the depths between 0.2 and 5.0 meters.

KITTI dataset [49] depth completion benchmark contains $\approx$86,000 raw $1242\times$ 375 image frames (43K stereo pairs) and synchronized sparse depth maps . The

Table 2: **Different Teacher Ensembles.** We apply Monitored Distillation to various combinations of teachers trained on different datasets. Using an ensemble trained only on NYUv2(‡) and SceneNet(†) still benefits a student on VOID(⋄).

Teachers	Teachers Trained On	MAE	RMSE	iMAE	iRMSE
None (Unsupervised Loss Only)	-	55.67	117.21	28.68	58.31
FusionNet$^{\diamond}$, ScaffNet†	VOID, SceneNet	48.72	102.44	26.94	56.32
FusionNet$^{\diamond}$, KBNet$^{\diamond}$	VOID	40.10	92.03	22.16	46.86
KBNet$^{\diamond}$, ScaffNet†	VOID, SceneNet	38.87	91.76	20.50	46.67
FusionNet$^{\diamond}$, KBNet$^{\diamond}$, ScaffNet†	VOID, SceneNet	36.42	87.78	19.18	43.83
FusionNet‡, KBNet‡, ScaffNet†	NYUv2, SceneNet	46.66	104.05	26.13	54.96

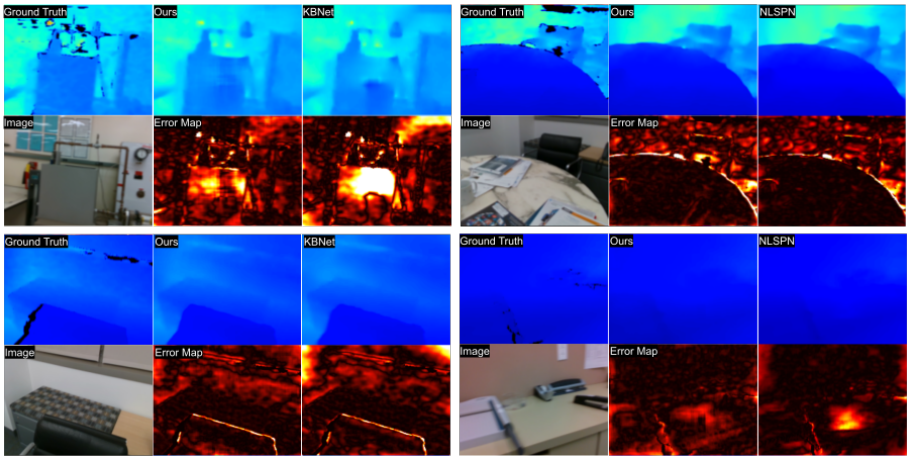


Fig. 4: **Monitored Distillation vs. Unsupervised (Left) and Supervised (Right) Teachers**. While KBNet [58] is the best performer among unsupervised methods, by ensembling it with weaker methods, we addressed its error modes on lab equipment (top left). Similarly, we address the failure modes in the top supervised method NLSPN [39] by distilling from a heterogeneous ensemble

sparse depth is obtained using a Velodyne lidar sensor and, when projected, covers $\approx 5\%$ of the image space. The ground truth depths are semi-dense, which we use only for evaluation purposes. We use the designated 1,000 samples for validation and evaluate test-time accuracy on KITTI's online testing server.

Teacher ensembles: We use the following ensembles for VOID (Table 1, 2, 3): (i) supervised ensemble of NLSPN [39], MSG-CHN [29], ENet, and PENet [21], (ii) unsupervised ensemble of FusionNet [54], KBNet [58], and ScaffNet [54] (trained on SceneNet [36]), and (iii) heterogeneous ensemble of all seven methods. For KITTI (Table 4, 5), we used NLSPN [39], ENet, and PENet [21].

VOID Depth Completion Benchmark. We present qualitative and quantitative experiments on VOID against blind ensemble distillation baselines, and top

Table 3: **VOID Benchmark.** We compare against unsupervised (U) and supervised (S) methods. By distilling from blind ensemble (BE), we outperform all existing works except for [39] which has 5× more parameters. Using our method with an unsupervised ensemble also yields 1st among unsupervised methods.

Method	Type	# Param	Time	MAE	RMSE	iMAE	iRMSE
SS-S2D [35]	U	27.8M	59ms	178.85	243.84	80.12	107.69
DDP [62]	U	18.8M	54ms	151.86	222.36	74.59	112.36
VOICED [56]	U	9.7M	29ms	85.05	169.79	48.92	104.02
ScaffNet [54]	U	7.8M	25ms	59.53	119.14	35.72	68.36
ENet [21]	S	131.7M	75ms	46.90	94.35	26.78	52.58
MSG-CHN [29]	S	364K	36ms	43.57	109.94	23.44	52.09
KBNet [58]	U	6.9M	13ms	39.80	95.86	21.16	49.72
Ours (Unsupervised)	**BE**	**5.3M**	**13ms**	**36.42**	**87.78**	**19.18**	**43.83**
PENet [21]	S	132M	226ms	34.61	82.01	18.89	40.36
Ours (Supervised)	**BE**	**5.3M**	**13ms**	**30.88**	**87.48**	**15.31**	**38.33**
Ours (Heterogeneous)	**BE**	**5.3M**	**13ms**	**29.67**	**79.78**	**14.84**	**37.88**
NLSPN [39]	*S*	*25.8M*	*122ms*	*26.74*	*79.12*	*12.70*	*33.88*

supervised and unsupervised methods. Note that while we evaluate our method and baselines across different ensemble compositions, Monitored Distillation and baselines have no knowledge regarding any individual teacher in the ensemble. For comparison purposes, scores for each teacher can be found in Table 3.

Comparisons Against Baselines: As we are the first to propose knowledge distillation for blind ensembles (Table 1), we begin by presenting several baselines: (1) mean, and (2) median of teachers, and (3) randomly selecting a teacher for each sample per iteration. All baselines are trained with distillation and unsupervised loss, unless specified otherwise, for fair comparisons against our method – which also consistently improves results for all ensemble types.

Table 1 row 1 shows the baseline performance of the student network trained only on unsupervised losses. Compared to the KBNet [58] in Table 3 row 7, our compressed KBNet (student) has a 23.2% sharp drop in performance due to a decrease in capacity. While all distillation methods improves its performance, Monitored Distillation beats all baselines by an average of 8.53% when using an ensemble of supervised teachers. This improvement grows to 11.50% when using an unsupervised ensemble (Table 1, rows 8-13), where the variance in teacher performance is considerably higher than supervised ones. Nonetheless, distilling an unsupervised ensemble improves over the best unsupervised method KBNet by an average of 9.53% – showing that we can indeed leverage the strengths of "weaker" methods to address the weakness of even the best method.

When using our method to distill from a heterogeneous ensemble, we observe the same trend where adding more teachers produces a stronger overall ensemble – improving over both supervised and unsupervised ones alone. This is unlike

Table 4: **KITTI Unsupervised Depth Completion Benchmark**. Our method outperforms all unsupervised methods across all metrics on the KITTI leaderboard. * denotes methods that use additional synthetic data for training.

Method	# Param	Time	MAE	RMSE	iMAE	iRMSE
SS-S2D [35]	27.8M	80ms	350.32	1299.85	1.57	4.07
IP-Basic [27]	0	11ms	302.60	1288.46	1.29	3.78
DFuseNet [46]	n/a	80ms	429.93	1206.66	1.79	3.62
DDP* [62]	18.8M	80ms	343.46	1263.19	1.32	3.58
VOICED [56]	9.7M	44ms	299.41	1169.97	1.20	3.56
AdaFrame [55]	6.4M	40ms	291.62	1125.67	1.16	3.32
SynthProj* [34]	2.6M	60ms	280.42	1095.26	1.19	3.53
ScaffNet* [54]	7.8M	32ms	280.76	1121.93	1.15	3.30
KBNet [58]	6.9M	16ms	256.76	1069.47	1.02	2.95
Ours	**5.3M**	**16ms**	**218.60**	**785.06**	**0.92**	**2.11**

naive distillation baselines, where "polluting" the ensemble with weaker teachers results in a drop in performance (Table 1, rows 14-19). In fact, naive distillation of heterogeneous ensemble is only marginally better than distilling unsupervised ensemble and considerably worse than a supervised one. In contrast, our method improves over baselines across all metrics by average of 17.53% and by 4.60% and 16.28% over our method with supervised and unsupervised ensembles, respectively. We also show an ablation study for β (Eqn. 4) by removing sparse depth error from our validation criterion, where we observe an average drop of 4.32% without β across all ensemble types due to the inherent ambiguity in scale when using monocular images for reconstruction. β allows us to choose not only predictions that yield high fidelity reconstructions, but also metric scale.

Different Teacher Ensembles: Table 2 shows the effect of having different teacher combinations within the ensemble. In general, the more teachers the better, and the better the teachers, the better the student. For example, combinations of any two teachers from the unsupervised ensemble yields less a performant student than the full ensemble of FusionNet, KBNet and ScaffNet – including that adding an underperforming method like ScaffNet to the ensemble (rows 3, 5). Finally, we show in row 6 that distilling from an ensemble trained on completely different datasets than the target test dataset (i.e. KBNet and FusionNet are trained on NYU v2 [47] and Scaffnet on SceneNet [36]) still improves over unsupervised loss with generic regularizers like local smoothness (row 1).

Benchmark Comparisons: Table 3 shows comparisons on the VOID benchmark. In an indoor setting, scene layouts are very complex with point clouds typically in orders of hundreds to several thousand points. As such, there are many suitable dense representations that can complete a given point cloud. Hence, the accuracy of the model hinges on the regularization as most of the scene does

Table 5: **KITTI Supervised Depth Completion Benchmark.** We compare against distilled (D) and supervised (S) methods. Despite operating in the blind ensemble (BE) distillation regime, our method beats many supervised methods. Our iMAE (0.92) and iRMSE (2.11) scores rank 4th, and we tie for 5th overall. Note: a method outranks another if it performs better on more than two metrics.

Rank	Method	Type	MAE	RMSE	iMAE	iRMSE
13	CSPN [7]	S	279.46	1019.64	1.15	2.93
12	SS-S2D [35]	S	249.95	814.73	1.21	2.80
9	Self-Distill [23]	D	248.22	949.85	0.98	2.48
9	DeepLiDAR [42]	S	226.50	758.38	1.15	2.56
9	PwP [60]	S	235.73	785.57	1.07	2.52
8	UberATG-FuseNet [5]	S	221.19	752.88	1.14	2.34
5	**Ours**	**BE**	**218.60**	**785.06**	**0.92**	**2.11**
5	RGB_guide&certainty [50]	S	215.02	772.87	0.93	2.19
5	ENet [21]	S	216.26	741.30	0.95	2.14
4	PENet [21]	S	210.55	730.08	0.94	2.17
2	DDP [62]	S	203.96	832.94	0.85	2.10
2	CSPN++ [6]	S	209.28	743.69	0.90	2.07
1	*NLSPN [39]*	*S*	*199.59*	*741.68*	*0.84*	*1.99*

not allow for establishing unique correspondences due to largely homogeneous regions, occlusions and the aperture problem.

Unlike generic regularizers (e.g. piecewise-smoothness), Monitored Distillation is informed by the statistics of many other scenes. Hence, even when distilling from an unsupervised ensemble (row 8), we still beat the best unsupervised method, KBNet [58], by an average of 9.53% over all metrics while using a 23.2% smaller model. This highlights the benefit of our positive congruent training, where our distillation objective can address the error modes of individual teachers. This is shown in Fig. 4 (top left), where we fixed the erroneous predictions in the lab equipment, and reduced errors in homogeneous regions.

Furthermore, distilling from a heterogeneous ensemble yields a student that ranks 2nd on the benchmark, achieving comparable performance to the top method NLSPN [39] while boosting a 79% model size reduction. Note: we do not outperform NLSPN despite it being included in the ensemble. This is likely due to distillation loss from the large size reduction. Fig. 4 shows that our model distills complex priors from the teacher ensemble regarding the topology of indoor surfaces, e.g. cabinets and tables are flat. The bottom right of Fig. 4 shows that we can even produce a more accurate depth estimate than NLSPN.

KITTI Depth Completion Benchmark. We provide quantitative comparisons against unsupervised and supervised methods on the KITTI test set. We also provide qualitative comparisons in Supp. Mat.

Comparison with Unsupervised Methods: Table 4 shows that despite having fewer parameters than most unsupervised models (e.g. 23.2% fewer than KBNet[58], 73.0% fewer than DDP[62]), our method outperforms the state of the art [58] across all metrics by an average of 19.93%, and by as much as 28.47% in iRMSE while boasting a 16ms inference time. Compared to methods that use synthetic ground truth to obtain a learned prior (marked with * in Table 4), our method leverages learned priors from pretrained models and improves over [34,54] by an average of 28.32% and 27.06%. We posit that this is largely due to the sim2real domain gap that [34,54,62] have to overcome i.e. covariate shift due to image translation error during training.

Comparison with Distilled and Supervised Methods: We compare our method (having at best *indirect* access to ground truth) against supervised and distilled methods that have *direct* access to ground truth in training. Table 5 shows that we rank 4th in iMAE and iRMSE, and tie for 5th overall. Note: We beat knowledge distillation method Self-Distill [23] by 12.6% despite (i) they use ground truth and (ii) we apply our method in the blind ensemble setting. We achieve comparable performance to the teacher models ENet [21] (131.7M params), PENet [21] (132M params), and NLSPN [39] (25.8M params) across all metrics despite only requiring 5.3M parameters.

5 Discussion

We propose Monitored Distillation for blind ensemble learning and knowledge distillation on depth completion tasks. Our method is capable of shrinking model size by 79% compared to the best teacher model, while still attaining comparable performance, enabling lightweight and deployable models.

However, we note that there exists several *risks and limitations.* (i) Our method relies on the composition of teachers and their error modes; if all teachers perform poorly on certain regions, our performance in these regions will not improve beyond training with unsupervised losses. (ii) Our method relies on structure-from-motion. If there is insufficient parallax between the stereo or monocular images, then photometric reprojection is uninformative regarding the depth of the scene. (iii) Reprojection error is limited when Lambertian assumptions are violated. However, the domain coverage of specularities and translucency is sparse due to the sparsity of primary illuminants [25] (rank of the reflectance tensor is deficient and typically small). So, explicitly modeling deviations from diffuse Lambertian reflection is likely to yield modest returns.

Admittedly the scope of this work is limited to depth completion, but we foresee this method being applied to general geometric problems (e.g. optical flow, stereo). Our method is the first attempt in blind ensemble distillation to produce positive congruent students, and we hope it lays the groundwork for approaches aiming to ensemble the abundance of existing pretrained models.

Acknowledgements. This work was supported by ARO W911NF-17-1-0304, ONR N00014-22-1-2252, NIH-NEI 1R01EY030595, and IITP-2021-0-01341 (AIGS-CAU). We thank Stefano Soatto for his continued support.

References

1. Chao, C.H., Cheng, B.W., Lee, C.Y.: Rethinking ensemble-distillation for semantic segmentation based unsupervised domain adaption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2610–2620 (2021)
2. Chawla, A., Yin, H., Molchanov, P., Alvarez, J.: Data-free knowledge distillation for object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3289–3298 (2021)
3. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems **30** (2017)
4. Chen, L., Yu, C., Chen, L.: A new knowledge distillation for incremental object detection. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2019)
5. Chen, Y., Yang, B., Liang, M., Urtasun, R.: Learning joint 2d-3d representations for depth completion. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10023–10032 (2019)
6. Cheng, X., Wang, P., Guan, C., Yang, R.: Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10615–10622 (2020)
7. Cheng, X., Wang, P., Yang, R.: Depth estimation via affinity learned with convolutional spatial propagation network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 103–119 (2018)
8. Chodosh, N., Wang, C., Lucey, S.: Deep convolutional compressed sensing for lidar depth completion. In: Asian Conference on Computer Vision. pp. 499–513. Springer (2018)
9. Choi, K., Jeong, S., Kim, Y., Sohn, K.: Stereo-augmented depth completion from a single rgb-lidar image. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13641–13647. IEEE (2021)
10. Dimitrievski, M., Veelaert, P., Philips, W.: Learning morphological operators for depth completion. In: International Conference on Advanced Concepts for Intelligent Vision Systems. Springer (2018)
11. Eldesokey, A., Felsberg, M., Holmquist, K., Persson, M.: Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12014–12023 (2020)
12. Eldesokey, A., Felsberg, M., Khan, F.S.: Propagating confidences through cnns for sparse data regression. In: Proceedings of British Machine Vision Conference (BMVC) (2018)
13. Fei, X., Wong, A., Soatto, S.: Geo-supervised visual depth prediction. IEEE Robotics and Automation Letters **4**(2), 1661–1668 (2019)
14. Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J., Ramabhadran, B.: Efficient knowledge distillation from an ensemble of teachers. In: Interspeech. pp. 3697–3701 (2017)
15. Gofer, E., Praisler, S., Gilboa, G.: Adaptive lidar sampling and depth completion using ensemble variance. IEEE Transactions on Image Processing (2021)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

17. Hong, B.W., Koo, J.K., Dirks, H., Burger, M.: Adaptive regularization in convex composite optimization for variational imaging problems. In: German Conference on Pattern Recognition. pp. 268–280. Springer (2017)
18. Hong, B.W., Koo, J., Burger, M., Soatto, S.: Adaptive regularization of some inverse problems in image analysis. IEEE Transactions on Image Processing (2019)
19. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
20. Hu, J., Fan, C., Jiang, H., Guo, X., Gao, Y., Lu, X., Lam, T.L.: Boosting light-weight depth estimation via knowledge distillation. arXiv preprint arXiv:2105.06143 (2021)
21. Hu, M., Wang, S., Li, B., Ning, S., Fan, L., Gong, X.: Penet: Towards precise and efficient image guided depth completion. arXiv preprint arXiv:2103.00783 (2021)
22. Huang, Z., Fan, J., Cheng, S., Yi, S., Wang, X., Li, H.: Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion. IEEE Transactions on Image Processing **29**, 3429–3441 (2019)
23. Hwang, S., Lee, J., Kim, W.J., Woo, S., Lee, K., Lee, S.: Lidar depth completion using color-embedded information via knowledge distillation. IEEE Transactions on Intelligent Transportation Systems (2021)
24. Jaritz, M., De Charette, R., Wirbel, E., Perrotton, X., Nashashibi, F.: Sparse and dense data with cnns: Depth completion and semantic segmentation. In: 2018 International Conference on 3D Vision (3DV). pp. 52–60. IEEE (2018)
25. Jin, H., Soatto, S., Yezzi, A.J.: Multi-view stereo beyond lambert. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 1, pp. I–I. IEEE (2003)
26. Kang, J., Gwak, J.: Ensemble learning of lightweight deep learning models using knowledge distillation for image classification. Mathematics **8**(10), 1652 (2020)
27. Ku, J., Harakeh, A., Waslander, S.L.: In defense of classical image processing: Fast depth completion on the cpu. In: 2018 15th Conference on Computer and Robot Vision (CRV). pp. 16–22. IEEE (2018)
28. Lan, X., Zhu, X., Gong, S.: Knowledge distillation by on-the-fly native ensemble. arXiv preprint arXiv:1806.04606 (2018)
29. Li, A., Yuan, Z., Ling, Y., Chi, W., Zhang, C., et al.: A multi-scale guided cascade hourglass network for depth completion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 32–40 (2020)
30. Liu, T.Y., Agrawal, P., Chen, A., Hong, B.W., Wong, A.: Monitored distillation for positive congruent depth completion. arXiv preprint arXiv:2203.16034 (2022)
31. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2604–2613 (2019)
32. Liu, Y., Shu, C., Wang, J., Shen, C.: Structured knowledge distillation for dense prediction. IEEE transactions on pattern analysis and machine intelligence (2020)
33. Liu, Y., Sheng, L., Shao, J., Yan, J., Xiang, S., Pan, C.: Multi-label image classification via knowledge distillation from weakly-supervised detection. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 700–708 (2018)
34. Lopez-Rodriguez, A., Busam, B., Mikolajczyk, K.: Project to adapt: Domain adaptation for depth completion from noisy and sparse sensor data. In: Proceedings of the Asian Conference on Computer Vision (2020)
35. Ma, F., Cavalheiro, G.V., Karaman, S.: Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In: International Conference on Robotics and Automation (ICRA). pp. 3288–3295. IEEE (2019)

36. McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2678–2687 (2017)
37. Merrill, N., Geneva, P., Huang, G.: Robust monocular visual-inertial depth completion for embedded systems. In: International Conference on Robotics and Automation (ICRA). IEEE (2021)
38. Michieli, U., Zanuttigh, P.: Knowledge distillation for incremental learning in semantic segmentation. Computer Vision and Image Understanding **205**, 103167 (2021)
39. Park, J., Joo, K., Hu, Z., Liu, C.K., Kweon, I.S.: Non-local spatial propagation network for depth completion. In: European Conference on Computer Vision, ECCV 2020. European Conference on Computer Vision (2020)
40. Park, S., Heo, Y.S.: Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. Sensors **20**(16), 4616 (2020)
41. Pilzer, A., Lathuiliere, S., Sebe, N., Ricci, E.: Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9768–9777 (2019)
42. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3313–3322 (2019)
43. Qu, C., Liu, W., Taylor, C.J.: Bayesian deep basis fitting for depth completion with uncertainty. arXiv preprint arXiv:2103.15254 (2021)
44. Qu, C., Nguyen, T., Taylor, C.: Depth completion via deep basis fitting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 71–80 (2020)
45. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
46. Shivakumar, S.S., Nguyen, T., Miller, I.D., Chen, S.W., Kumar, V., Taylor, C.J.: Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 13–20. IEEE (2019)
47. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
48. Traganitis, P.A., Giannakis, G.B.: Blind multi-class ensemble learning with dependent classifiers. In: 2018 26th European Signal Processing Conference (EUSIPCO). pp. 2025–2029. IEEE (2018)
49. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: 2017 International Conference on 3D Vision (3DV). pp. 11–20. IEEE (2017)
50. Van Gansbeke, W., Neven, D., De Brabandere, B., Van Gool, L.: Sparse and noisy lidar completion with rgb guidance and uncertainty. In: 2019 16th International Conference on Machine Vision Applications (MVA). pp. 1–6. IEEE (2019)
51. Walawalkar, D., Shen, Z., Savvides, M.: Online ensemble model compression using knowledge distillation. In: European Conference on Computer Vision. pp. 18–35. Springer (2020)

52. Wang, Y., Li, X., Shi, M., Xian, K., Cao, Z.: Knowledge distillation for fast and accurate monocular depth estimation on mobile devices. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2457–2465 (2021)
53. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
54. Wong, A., Cicek, S., Soatto, S.: Learning topology from synthetic data for unsupervised depth completion. IEEE Robotics and Automation Letters **6**(2), 1495–1502 (2021)
55. Wong, A., Fei, X., Hong, B.W., Soatto, S.: An adaptive framework for learning unsupervised depth completion. IEEE Robotics and Automation Letters **6**(2), 3120–3127 (2021)
56. Wong, A., Fei, X., Tsuei, S., Soatto, S.: Unsupervised depth completion from visual inertial odometry. IEEE Robotics and Automation Letters (2020)
57. Wong, A., Soatto, S.: Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5644–5653 (2019)
58. Wong, A., Soatto, S.: Unsupervised depth completion with calibrated backprojection layers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12747–12756 (2021)
59. Xiang, L., Ding, G., Han, J.: Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In: European Conference on Computer Vision. pp. 247–263. Springer (2020)
60. Xu, Y., Zhu, X., Shi, J., Zhang, G., Bao, H., Li, H.: Depth completion from sparse lidar data with depth-normal constraints. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2811–2820 (2019)
61. Yan, S., Xiong, Y., Kundu, K., Yang, S., Deng, S., Wang, M., Xia, W., Soatto, S.: Positive-congruent training: Towards regression-free model updates. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14299–14308 (2021)
62. Yang, Y., Wong, A., Soatto, S.: Dense depth posterior (ddp) from single image and sparse range. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3353–3362 (2019)
63. Zhang, Y., Funkhouser, T.: Deep depth completion of a single rgb-d image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 175–185 (2018)