AugUndo: Scaling Up Augmentations for Monocular Depth Completion and Estimation

Supplementary Material

Summary. In Sec. A, we provide implementation details and hyperparameters, such as learning rate schedule and crop size, used to reproduce our results. In B, we provide augmentation parameters used to train each method on each dataset. Sec. C lists the evaluation metrics used in our quantitative results. To illustrate the motivation behind AugUndo, we provide examples of image and sparse depth degradation through typical augmentation routines and add an extended discussion in Sec. D. We provide details on datasets used in our experiments in Sec. E. In Sec. F, we provide a detail walk-through of the AugUndo algorithm. To study the contribution of each of the proposed augmentation, we present a comprehensive ablation study in Sec. G. We further extend the sensitivity study to sparsity from the main paper conducted on VOID to VOID500 and VOID150 in Sec. H. We present our results on zero-shot generalization from KITTI to Waymo (depth completion) and Make3d (monocular depth estimation) in Sec. I. We also show a sensitivity study on the effect of different augmentations in Sec. J. In Sec. K, we include additional results for depth completion, including extensive quantitative results such as evaluations on MonDi and DesNet, comparing modeling AugUndo as change in camera pose and parameters, and the effects of naively applying geometric augmentations during unsupervised training. In Sec. L, we show additional results on monocular depth estimation. Finally, we discuss limitations in Sec. M.

A Implementation details

For unsupervised depth completion, we implemented our method in PyTorch and incorporated our augmentation pipeline into the codebases of VOICED [45], FusionNet [43], KBNet [48], MonDi [22], and DesNet [50]. For monocular depth estimation, we implement our method in PyTorch according to [11]. Specifically, we implemented our augmentation pipeline into the codebases of Monodepth2 [11], HR-Depth [23], and Lite-Mono [55]. Details of each task are described below.

Unsupervised depth completion. The models are optimized using Adam [14] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For VOID: We used an input batch size of 12 with random crop size of 416×512 for KBNet and DesNet, a batch size of 8 without cropping for FusionNet and VOICED, and a batch size of 8 with random crop size of 448×576 for MonDi. We trained each KBNet and DesNet for 40 epochs with base learning rate of 1×10^{-4} for 20 epochs and decreased to 5×10^{-5} for the last 20 epochs. We trained FusionNet and VOICED for 20 epochs with a base learning rate of 1×10^{-4} for 10 epochs and decreased to 5×10^{-5} for the last 10 epochs. For KITTI: We used a batch size of 8 and random crop size of 320 by 768 for all models. We trained KBNet for 60 epochs, with 5×10^{-5} for

2 epochs, 1×10^{-4} until 8th epoch, 2×10^{-4} until 30th epoch, 1×10^{-4} until the 45th epoch and 5×10^{-5} until the 60th epoch. We trained FusionNet for 30 epochs 2×10^{-4} for 16 epochs, 1×10^{-4} until 24th epoch and 5×10^{-5} until the 30th epoch. We trained VOICED for 30 epochs 2×10^{-4} for 16 epochs, 6×10^{-5} until 24th epoch and 3×10^{-5} until the 30th epoch.

We ensure that all baseline methods can reproduce or exceed the numbers originally reported by the authors. Since the authors of DesNet did not publish their code implementation, we re-implemented their method to the best of our ability. All reported results of AugUndo are based on those same settings with the exception of the augmentation scheme.

Unsupervised monocular depth estimation. The models are optimized using Adam [14] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For VOID, we used a input batch size of 12 and random crop size of 256×448 for all models. For Monodepth2, we trained with learning rate 1×10^{-8} for 2 epochs, 1×10^{-5} for the next 23 epochs, and 1×10^{-6} for the next 25 epochs. For HR-Depth, we train with learning rate 1^{-4} for 15 epochs, and 1×10^{-5} for the final 5 epochs. For Lite-Mono, we train with learning rate 5×10^{-4} for 35 epochs. For KITTI dataset, We used a input batch size of 12 and a random crop size of 192×640. We trained the models for 20 epochs with an initial learning rate of 1×10^{-4} and drop the learning rate to 1×10^{-5} at the 15th epoch. The smoothness loss weight for KITTI is set to 0.001 as per [11] and 0.01 for VOID as VOID contains more indoor scenes with homogeneous surfaces.

We ensure that all baseline methods can reproduce or exceed the numbers originally reported by the authors. All reported results of AugUndo are based on those same settings with the exception of the augmentation scheme. Note: For Monodepth2 and HR-Depth, we initialize the ResNet encoder weight with the Imagenet-pretrained weight downloaded from PyTorch website, as specified in their Github repository. However, we cannot locate the pretrained weight used for Lite-Mono throughout their repository, which left us no choice but to train their model from scratch. Nonetheless, this does not affect the validity of the study as the comparison is made between models initialized from scratch where the difference is only in the augmentations schemes: standard convention used in the original papers and repositories, and AugUndo.

B Augmentations

Through a search over augmentation types and values, we found a consistent set of augmentations that tends to yield improvements across all methods with small changes to degree of augmentation catered to each method. All augmentations are applied with a 50% probability. We note that performance gain can be obtained by typical set and ranges of augmentations and does not require a meticulous selection of hyper-parameters (see Fig. 2).

VOID. For depth completion, we applied photometric transformations of random brightness from 0.5 to 1.5, contrast from 0.5 to 1.5, saturation from 0.5 to 1.5, and hue from -0.1 to 0.1. We applied image patch removal by selecting

between 0.1% to 0.5% of the pixels and removing 5 × 5 patches centered on them – approximately removing between 2.5% to 12.5% of the image. We applied random sparse depth point removal at a rate between 60% and 70% of all sparse points. We applied geometric transformations of random horizontal and vertical flips, up to 10% of translation and between -25 to 25 degrees of rotation. For KBNet and FusionNet, we applied random resize factor between 0.6 to 1.1, while for VOICED, we used 0.7 to 1.1 because the scaffolding step of VOICED requires at least 3 points, but as discussed above, resampling causes loss and factors smaller than 0.7 often cause all points to be dropped. For monocular depth estimation, we applied photometric transformations of random brightness from 0.5 to 1.5, contrast from 0.5 to 1.5, saturation from 0.5 to 1.5, hue from -0.1 to 0.1. We applied geometric transformation of random rotation between -10 to 10 degrees and random horizontal flipping. We further applied random resize factor between 0.8 to 1.

KITTI. For depth completion, we applied random brightness, contrast, saturation from 0.5 to 1.5 and random hue from -0.1 to 0.1. We applied image patch removal by selecting between 0.1% to 0.5% of the pixels and removing 5 × 5 patches centered on them. We applied random sparse depth point removal at a rate between 60% and 70% of all sparse points. We further applied random horizontal flips, up to 10% of translation, resizing factors between 0.8 to 1.2, and between -20 to 20 degrees of rotation. We found that vertical flips are detrimental to performance. For monocular depth estimation, we applied random brightness, contrast, saturation from 0.5 to 1.5 and random hue from -0.1 to 0.1. We applied a random rotation between -30 to 30 degrees and a random translation of up to 30% of the image. We further apply horizontal flip to the image.

Table 1: *Error metrics for depth completion and monocular depth estimation.* d_{gt} denotes ground truth depth and evaluated where values are available for a given image.

Metric	Definition
MAE	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d_{gt}(x)\rvert$
RMSE	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d_{gt}(x)\rvert^2\right)^{1/2}$
iMAE	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{d}(x)-1/d_{gt}(x)\rvert$
iRMSE	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{d}(x)-1/d_{gt}(x)\rvert^2\right)^{1/2}$
AbsRel	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\frac{\lvert\hat{d}(x)-d_{gt}(x)\rvert}{d_{gt}(x)}$
SqRel	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\frac{\lvert\hat{d}(x)-d_{gt}(x)\rvert^2}{d_{gt}(x)}$
Accuracy % of $z(x)$ s.t. $\delta \doteq \max(\frac{z(x)}{z_{gt}(x)}, \frac{z_{gt}(x)}{z(x)}) <$ threshold	

C Evaluation metrics

The evaluation metrics used for depth completion and monocular depth estimation are shown in Tab. 1. Depth completion models are evaluated with MAE,

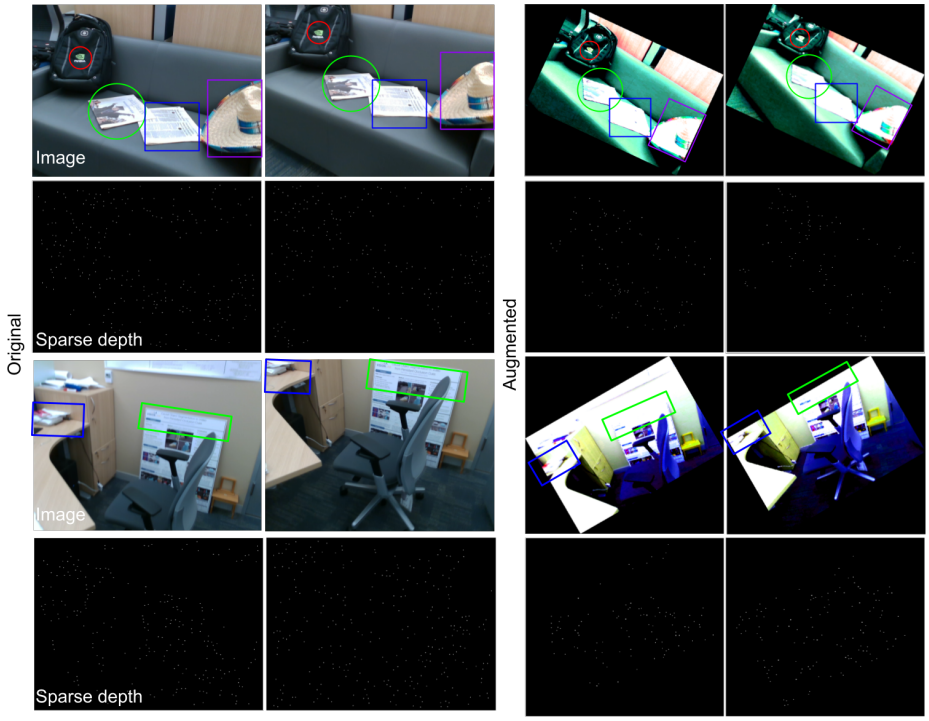

Fig. 1: *Motivation.* After augmenting the inputs with photometric and geometric transformations, the previously exciting image textures (for establishing correspondence across images) and sparse points in the original inputs, that would have served as supervision, are now largely saturated in intensity and homogeneous, and lost due to resampling, respectively (best viewed in 2×).

RMSE, iMAE, iRMSE. Monocular depth estimation models are evaluated with MAE, RMSE, AbsRel, SqRel, and accuracy ($\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$). We note that as monocular depth is inferred, at most, up to an unknown scale, we perform scale matching during evaluation by using median scaling with respect to the ground truth before computing each error or accuracy metric.

D Extended Discussion of Motivation

In the main paper, we discussed the motivation behind the approach. Here, we provide an extended discussion: Training unsupervised depth completion methods [43–45, 48, 52] relies on a photometric reconstruction term, sparse depth reconstruction term, and a (generic) regularizer, such as local smoothness; a related problem, unsupervised monocular depth estimation [8,11,20,42,47,54], omits the sparse depth term. The photometric reconstruction term constraints the solution in regions where there are sufficiently exciting textures and co-visible between images such as those highlighted in Fig. 1. The sparse depth reconstruction term

constraints the solution anywhere with a sparse point. Everywhere else in the image are inherently ambiguous, so we must rely on the regularizer. Fig. 1 shows that after one has applied a number of photometric and geometric transformations, the previously exciting textures that would have served as our supervision are now largely saturated in intensity and homogenous. The number of points in the sparse depth maps have also been greatly reduced.

In conventional augmentation scheme for *other* (semantic) vision tasks, i.e. classification, detection, and segmentation, it is typical for one to introduce saturation in image intensities, block artifacts from interpolation and loss during resampling. These are some of the side-effects that are *desirable* in the aforementioned tasks: introducing these nuisance variabilities enables the model to learn to be invariant to them (i.e., to learn them away), which yields more generalizable and robust representations. This is also true for supervised depth completion [7, 25, 31, 39] and estimation methods [5, 26, 27, 38, 53]. However, this is not the case for learning unsupervised depth. Because the supervision signal comes from reconstructing the input image and sparse depth map, the more we augment the data (causing a loss of photometric correspondences across image frames and a loss of points in the sparse depth map), the more we degrade the supervision signal (see Tabs. 8 and 9 in Sec. K for empirical evidence of this phenomenon). So, it is not too surprising that conventional augmentation procedures, both photometric and geometric, have seen limited use beyond small changes in image intensities, and flipping.

Nevertheless, photometric augmentations help model the diverse range of illumination conditions and colors of object that may populate the scene. Geometric augmentations can simulate the various camera parameters and motion, for example, image translation can simulate principle point offsets or it can approximate small baseline movements, image resizing with cropping can simulate different focal lengths (zooming) or it can model camera movements, while in-plane rotations can model camera orientation. These augmentations are often viewed as essential to training pipelines for other vision tasks, but are detrimental for unsupervised depth completion; yet, without them, one may encounter robustness and generalization issues. As the root of the problem lies in the supervision signal, we investigate an approach to "undo" the augmentations during (or right before) the loss computation step. Doing so enables one to feed forward augmented inputs, but compute the loss on the original inputs with no loss in the training signal. Extensive experiments show that our approach improves performance and zero-shot generalization across a number of methods for both indoor and outdoor scenarios.

E Datasets

We conduct experiments on six datasets (two for training and four for generalization) in total. Details of each dataset are provided below.

KITTI [9] contains 61 driving scenes with research in autonomous driving and computer vision. It contains calibrated RGB images with sychronized point

clouds from Velodyne lidar, inertial, GPS information, etc. For depth completion [37], there are ≈80,000 raw image frames and associated sparse depth maps, each with a density of ≈5%. Ground-truth depth is obtained by accumulating 11 neighbouring raw lidar scans. Semi-dense depth is available for the lower 30% of the image space. We test on the official validation set of 1,000 samples because the online test server has submission restrictions to accomodate multiple trials. For depth estimation, we used Eigen split [5], following [58] to preprocess and remove static frames. The remaining training set contains 39,810 monocular triplets and the validation set contains 4,424 triplets. The testing set contains 697 monocular images. We follow the evaluation protocol of [6,37], where output depth is evaluated where ground truth exists between 0 to 80.0 meters.

VOID [45] comprises indoor (laboratories, classrooms) and outdoor (gardens) scenes with synchronized 640×480 RGB images and sparse depth maps. XIVO [8], a VIO system, is used to obtain the sparse depth maps that contain approximately 1500 sparse depth points with a density of about 0.5%. Active stereo is used to acquire the dense ground-truth depth maps. In contrast to the typically planar motion in KITTI, VOID has 56 sequences with challenging 6 DoF motion captured on rolling shutter. 48 sequences (about 45,000 frames) are assigned for training and 8 for testing (800 frames). We follow the evaluation protocol of [45] where output depth is evaluated where ground truth exists between 0.2 and 5.0 meters.

NYUv2 [30] consists of 372K synchronized 640×480 RGB images and depth maps for 464 indoors scenes (household, offices, commercial), captured with a Microsoft Kinect. The official split consisting in 249 training and 215 test scenes. We use the official test set of 654 images. Because there are no sparse depth maps provided, we sampled $\approx$ 1500 points from the depth map via Harris corner detector [13] to mimic the sparse depth produced by SLAM/VIO. We test models trained on VOID to evaluate their generalization to NYUv2. We follow the evaluation protocol of [45] where output depth is evaluated where ground truth exists between 0.2 and 5.0 meters.

ScanNet [4] consists of RGB-D data for 1,513 indoor scenes with 2.5 million images and corresponding dense depth map. Because there are no sparse depth maps provided, we sampled $\approx$ 1500 points from the depth map via Harris corner detector [13] to mimic the sparse depth produced by SLAM/VIO. We followed [4] and used 100 scenes (scene707-scene806), for zero-shot generalization for models trained on VOID. The output depth is evaluated where ground truth exists between 0.2 and 5.0 meters.

Waymo Open Dataset [34] contains 1920×1280 RGB images and lidar scans from autonomous vehicles. The training set contains ≈158K images from 798 scenes and the validation set ≈40K images from 202 scenes, collected at 10Hz. Objects are annotated across the full 360° field. Sparse depth maps are obtained by reprojecting the point cloud scan from the top lidar to the camera frame. Ground truth is obtained by reprojecting both front facing lidars as well as those collected 10 time steps forward and backwards (approximately 1 second of capture) to a given camera frame at a specific time step to densify the

Algorithm 1 AugUndo

Require: Depth completion network f_θ, Images I_t, I_τ, Sparse depth z_t, Relative pose g_{rt}, Intrinsics K

Sample $\{T^1_{pt,I} \dots T^k_{pt,I}\}$ from $T^i_{pt,I} \in \mathcal{A}_{pt,I}$, and compose $T_{pt,I} = T^1_{pt,I} \circ T^2_{pt,I} \circ \cdots \circ T^k_{pt,I}$

Sample $\{T^1_{pt,z} \dots T^j_{pt,z}\}$ from $T^i_{pt,z} \in \mathcal{A}_{pt,z}$, and compose $T_{pt,z} = T^1_{pt,z} \circ T^2_{pt,z} \circ \cdots \circ T^j_{pt,z}$

Sample $\{T^1_{ge} \dots T^m_{ge}\}$ from $T^i_{ge} \in \mathcal{A}_{ge}$, and compose $T_{ge} = T^1_{ge} \circ T^2_{ge} \circ \cdots \circ T^m_{ge}$

Compose the inverse geometric transform $T^{-1}_{ge} = (T^m_{ge})^{-1} \circ (T^{m-1}_{ge})^{-1} \circ \cdots \circ (T^1_{ge})^{-1}$

Compute the coordinates after geometric transform $\begin{bmatrix} x' & 1 \end{bmatrix}^\top = T_{ge} \begin{bmatrix} x & 1 \end{bmatrix}^\top$ (Eqn. 3 from main paper)

Augment I_t with photometric and geometric transformations $I'_t(x') = T_{pt,I}(I_t)(x)$ (Eqn. 4 from main paper)

Augment z_t with occlusion and geometric transformations $z'_t(x') = T_{pt,z}(z_t)(x)$ (Eqn. 4 from main paper)

Obtain depth prediction $\hat{d}'_t = f_\theta(I'_t, z'_t)$

Compute coordinates of the inverse geometric transformation $\begin{bmatrix} x'' & 1 \end{bmatrix}^\top = T^{-1}_{ge} \begin{bmatrix} x' & 1 \end{bmatrix}^\top$ (Eqn. 5 from main paper)

Apply inverse geometric transformation on output depth map: $\hat{d}_t(x'') = \hat{d}'_t(x')$ (Eqn. 6 from main paper)

Reconstruct I_t from I_τ using Eqn. 1 from main paper, i.e., $\hat{I}_{t\tau} = I_\tau(\pi g_{\tau t} K^{-1} \bar{x} \hat{d}_t)$

Minimize reconstruction losses between $\hat{I}_{t\tau}$ and I_t, and $\hat{d}_t$ and z_t, and the regularizer (Eqn. 2 from main paper)

sparse depth. We used the object annotations to remove all moving objects to ensure that reprojected points respects the static scene assumption. We also performed outlier removal to filter out errorenous (noisy) points. The output depth is evaluated where ground truth exists between a 1.5 and 80.0 meters range.

Make3d [29] contains 134 test images with 2272×1707 resolution. Ground-truth depth maps are given at 305×55 resolution and must be rescaled and interpolated. We use the central cropping proposed by [10] to get a 852×1707 center crop of the image. We use standard Make3d evaluation protocol and metrics. We use Make3d to test the generalization of monocular depth estimation models trained on KITTI.

F The AugUndo Algorithm

We assume that we are given (i) $I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3_+$ an RGB image I_t, (ii), its associated sparse point cloud projected onto as a depth map $z : \Omega_z \subset \Omega \to \mathbb{R}_+$, z_t, (iii) camera intrinsic calibration matrix $K \in \mathbb{R}^{3\times3}$, (iv) a sequence of images I_τ for $\tau \in \{t-1, t+1\}$ during training, and (v) the relative pose $g_{\tau t}$ between the image frame I_t and some temporally adjacent image I_τ. Given a image and its associated sparse depth map, a depth completion network learns a map

Table 2: *Ablation study for KBNet on VOID dataset.* BRI stands for brightness, CON for contrast, SAT for saturation, FLP for horizontal and vertical flip, TRN for translation, ROT for rotation, RZD for resize down, RZU for resize up, RMP for point removal, RMI for image patch removal. **Bold** denotes AugUndo, *italicized* the standard augmentation protocol. RMP, FLP, RZD, RZU have the highest influence; only when BRI, CON, SAT, HUE (color jitter) are disabled do they have non-negligible effect. Best results are achieved by using AugUndo.

Augmentation settings											Evaluation metrics			
BRI	CON	SAT	HUE	FLP	TRN	ROT	RZD	RZU	RMP	RMI	MAE	RMSE	iMAE	iRMSE
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**33.32**±0.18	**85.67**±0.39	**16.61**±0.29	**41.24**±0.60
✓	✓	✓	✓	✓							*38.11*±0.77	*95.22*±1.72	*19.51*±0.14	*46.70*±0.48
	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	33.46±0.27	86.02±0.69	16.84±0.22	41.78±0.63
✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	33.79±0.09	86.38±0.36	17.12±0.19	42.12±0.54
✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	33.73±0.30	85.84±0.22	17.05±0.16	41.73±0.47
✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	33.60±0.16	86.47±0.69	16.95±0.27	41.65±0.44
				✓	✓	✓	✓	✓	✓	✓	34.14±0.37	87.26±0.74	17.09±0.17	42.57±1.33
✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	44.38±0.89	106.87±0.89	22.91±0.68	52.55±0.57
✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	33.92±0.19	87.19±0.55	17.09±0.03	42.23±0.23
✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	33.68±0.19	85.86±0.51	16.92±0.03	41.47±0.18
✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	35.77±0.43	89.88±0.84	17.81±0.23	42.75±0.38
✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	35.69±0.25	90.40±0.71	17.75±0.12	42.82±0.20
✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	37.73±0.27	92.62±0.20	19.44±0.18	45.47±0.37
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		33.64±0.22	86.26±0.23	16.80±0.13	41.60±0.64

the inputs to the output depth map $\hat{d}_t := f_\theta(I_t, z_t) \in \mathbb{R}_+^{H \times W}$. In the main paper, we denoted photometric and occlusion augmentations as $\mathcal{A}_{pt}$ for ease of notation. Here, for specificity, we define $\mathcal{A}_{pt,I}$ as the set of possible photometric (including occlusion) transformations for the image, $\mathcal{A}_{pt,z}$ as the possible set of occlusion augmentations for sparse depth maps, and $\mathcal{A}_{ge}$ as the possible set of all geometric transformations. We additionally assumes we have the set of geometric transformations T_{ge} used during augmentation and their inverse transformations T_{ge}^{-1}. Alg. 1 is the procedural algorithm of AugUndo and details the step by step augmentation and loss computation pipelines, given our inputs.

G Ablation Study for Depth Completion

In the main paper, we demonstrated AugUndo for unsupervised depth completion methods on VOID1500 and VOID500 benchmarks using the most performant set of augmentations we found. Here, we provide an ablation study for each of the augmentations used to study their individual contributions.

Tab. 2 shows a comprehensive ablation study following augmentation settings above. From our best reported settings on KBNet, we removed individual augmentations to show their empirical contribution. Note: random brightness, contrast, saturation, and hue (BRI, SAT, CON, HUE) together is the standard color jitter employed by all current methods. We also test removing all color jitter augmentations to further quantify its effect.

We found that random flips (FLP) has the largest impact of all of the augmentations – increasing the error by an average of 30%; this is because it simulates

Table 3: *Sensitivity study of depth completion on VOID.* Reported scores are mean and standard deviation over four independent trials. We compare the sensitivity of models trained on VOID1500, with standard augmentations and AugUndo, by testing them on VOID500 and VOID150. Sparse depth maps within VOID1500 contains approximately 1500 points, and those within VOID500 and VOID150 contain approximately 3× and 10× less, respectively. AugUndo improves performance by an average of 19.66% and 19.62% across all unsupervised methods and evaluation metrics on VOID500 and VOID150, respectively. For distillation method (marked by *), MonDi, AugUndo improves by an average of 9.02% and 10.25% on VOID500 and VOID150, respectively. Despite MonDi is supervised by pseudo ground truth in addition to typical unsupervised losses, AugUndo can still boost performance.

	VOID500				VOID150			
Method	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓
VOICED	137.01±4.23	235.80±7.82	71.36±1.86	130.63±5.66	209.59±5.18	329.71±10.01	130.45±5.63	229.79±14.09
+ AugUndo	**92.99**±1.11	**176.94**±1.38	**46.43**±0.85	**91.10**±1.64	**151.77**±1.99	**262.61**±2.19	**88.65**±0.76	**169.29**±2.71
FusionNet	97.73±0.73	194.32±1.36	58.65±1.31	122.95±3.04	158.03±1.97	284.23±3.05	113.67±2.03	223.41±0.93
+ AugUndo	**74.97**±1.15	**162.71**±1.86	**40.44**±1.09	**92.11**±1.01	**126.16**±1.44	**246.16**±2.46	**86.13**±4.46	**181.08**±8.00
KBNet	78.44±1.39	178.17±3.27	37.56±0.61	83.43±1.89	149.13±3.29	306.30±8.74	70.74±2.26	136.75±5.55
+ AugUndo	**66.97**±0.81	**151.55**±2.03	**31.63**±0.53	**71.90**±0.82	**117.16**±4.51	**239.60**±10.96	**57.65**±1.47	**112.81**±1.75
DesNet	74.89±0.67	170.32±2.03	35.62±0.72	78.30±1.31	139.54±4.46	287.95±12.69	65.00±1.24	123.81±2.17
+ AugUndo	**67.78**±1.10	**153.46**±2.64	**32.09**±0.48	**71.96**±1.05	**117.93**±3.77	**239.49**±8.44	**58.13**±1.14	**112.78**±2.33
MonDi*	73.90±0.61	191.69±2.64	31.78±1.52	72.67±2.17	146.33±5.57	343.25±13.70	60.43±3.54	114.18±2.21
+ AugUndo	**69.90**±3.05	**157.02**±7.46	**29.61**±0.12	**68.48**±0.57	**127.08**±4.51	**299.42**±10.96	**54.47**±1.47	**108.23**±1.75

different scene layouts. As there are no viable intensity augmentations for sparse depth, RMP is the only one to explicitly increase variability in the data modality hence it also has large influence. We note that other geometric augmentations contribute as well, i.e., rotation, resizing, and translation, to discard points. By computing the loss on the original inputs, we reconstruct the removed points, which in fact serves as an additional training signal to map RGB to depth.

Photometric augmentations, individually, have small effect on performance. To get a non-neglible effect, one must disable all color jitter (Tab. 2 , row 7), which still yielded small increases in errors. This shows the limitations of existing augmentations, which relies heavily on color jittering. We note that we use a larger value range in color jitter than existing works, so the impact is expected to be even smaller for existing works. Admittedly, the most important one (FLP) is currently being used by all methods, but resize, image patch and point removal, play a large role. The best results are obtained when using all of the proposed augmentations, demonstrating the importance of scaling up both photometric and geometric augmentations.

H Sensitivity Study for Depth Completion

Given that conventional augmentation pipelines are not applied towards sparse depth modality, it is possible that a model will overfit to the sparse point cloud, which describes the coarse 3D scene structure. Overfitting to scenes, hence, can

limit generalization and increase sensitivity to the configuration of sparse points. In the main paper, we presented results for depth completion on VOID1500 and VOID500. Here, we test the effect of AugUndo on various sparse depth input densities. We presented results on VOID500 (repeated for side-by-side comparison) and VOID150 in Tab. 3 (left and right, respectively). VOID500 contains approximately 500 sparse points per point cloud and VOID150 contains approximately 150 points. All models tested are trained on VOID1500, which contains 1500 sparse points.

For VOID500 (Tab. 3, left), which is a 3× reduction in density, AugUndo improves the sensitivity to changes in the sparse points by an average of 19.66% across all methods, and 30.57%, 23.92%, 14.79%, 9.35% for VOICED, FusionNet, KBNet, and DesNet, respectively. For an even more challenging scenario, also the closest setup to the density of sparse points tracked by a Simultaneous Localization and Mapping (SLAM) and Visual Inertial Odometry (VIO) system, we consider VOID150 (Tab. 3, right). Here, AugUndo improves by an average of 19.62% across all methods, and 26.57%, 19.18%, 19.80%, and 12.95% for VOICED, FusionNet, KBNet, and DesNet, respectively. We attribute this to geometric and occlusion augmentations: translation, resize, rotation and sparse points removal. All of these augmentations not only affect photometry, but also the sparse point cloud where points are dropped due to resampling or explicitly removed, and additionally point cloud orientation is also altered.

We note that AugUndo is also helpful for distillation methods like MonDi [22]. As mentioned in the main text, we conjecture that artifacts caused from transformation of a piece-wise smooth depth map (in the case of supervised or distillation methods) are less severe than those of image and sparse point clouds. Hence, we were expecting the gains for supervised or distillation methods to be small compared to unsupervised methods. However, as shown in the last row of Tab. 3, we observe a surprisingly nontrivial gain when applying AugUndo to MonDi (marked by *). For VOID500, we improve MonDi by 9.02% on average across on metrics; for VOID150, we improve by 10.25%.

I Zero-shot Generalization from KITTI

In the main paper, we demonstrated AugUndo for three unsupervised depth completion (VOICED [45], FusionNet [43], and KBNet [48]) and three unsupervised monocular depth estimation (Monodepth2 [11], HR-Depth [23]), and Lite-Mono [55]) methods on the KITTI dataset. Due to space constraints, here, we provide additional results for zero-shot generalization from KITTI to Waymo Open Dataset [34] for depth completion and to Make3d [29] for monocular depth estimation. Similar to our generalization experiments on indoors (VOID to NYUv2 and ScanNet), we will train on KITTI using the conventional augmentation schemes employed by each respective method and compare the resulting models with those trained using AugUndo.

We begin by presenting results on depth completion in Tab. 4. Here, we evaluate VOICED, FusionNet, and KBNet models (trained on KITTI using standard

Table 4: *Zero-shot transfer from KITTI to Waymo for depth completion.* AugUndo improves generalization of models trained on KITTI to Waymo by an average of 13.3% for all evaluation metrics. We note that the sparse depth maps provided in Waymo is considerably denser than KITTI as they are merged from two front separate lidars; hence, FusionNet, which employs a learned (frozen) densification network performs similarly, i.e., the bias introduced by the same frozen densification network (ScaffNet) is strong enough that FusionNet yields similar results for both conventional augmentation scheme and AugUndo. Nonetheless, we still observe considerable improvements.

Dataset	Method	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓
Waymo	VOICED	6781.27±317.21	7734.57±339.77	24.58±1.45	27.87±1.63
	+ AugUndo	**5965.07**±367.47	**7029.66**±509.43	**18.54**±1.83	**22.10**±3.96
	FusionNet	530.55±39.23	1734.23±114.97	1.27±0.12	2.82±0.29
	+ AugUndo	**512.29**±8.43	**1707.34**±48.25	**1.21**±0.09	**2.75**±0.18
	KBNet	625.00±12.47	2167.74±92.17	1.76± 0.13	5.46±0.93
	+ AugUndo	**541.29**±15.16	**2014.14**±76.52	**1.34**±0.11	**3.43**±0.30

Table 5: *Zero-shot transfer from KITTI to Make3d for monocular depth estimation.* All models are trained on KITTI. Note: for Monodepth2, we use the numbers reported by [11] and the best trial on KITTI.

Dataset	Method	MAE ↓	RMSE ↓	Abs Rel ↓	Sq Rel ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Make3d	Monodepth2	-	7.417	0.322	3.589	-	-	-
	+ AugUndo	4.109	**6.803**	**0.272**	**2.769**	0.606	0.848	0.936
	HR-Depth	4.136	6.505	0.281	2.484	0.562	0.839	0.938
	+ AugUndo	**4.023**	**6.428**	**0.272**	**2.393**	**0.584**	**0.848**	**0.94**
	Lite-Mono	5.116	8.061	0.358	4.676	0.511	0.793	0.91
	+ AugUndo	**4.728**	**7.518**	**0.321**	**3.887**	**0.529**	**0.817**	**0.926**

augmentation pipelines and AugUndo) on the Waymo Open Dataset. Overall, training with AugUndo improves existing methods by an average 13.3% across all evaluation metrics. We note that AugUndo improves VOICED and KBNet by larger amounts than FusionNet. For FusionNet, whether trained with conventional augmentation scheme or AugUndo, both seem to perform similarly. This is because FusionNet employs a learning-based densification network (ScaffNet), which is pretrained on synthetic datasets and frozen, that is used in both models. As the sparse depth maps in Waymo are much denser than KITTI, ScaffNet is able to approximate the dense depth map with small amounts of errors. This serves as an inductive bias for the downstream FusionNet, which learns the residual over the approximated depth map. As the reconstruction from ScaffNet exhibits high fidelity, the bias induced by ScaffNet on FusionNet causes FusionNet to perform only minor modifications to approximated depth map, leading to similar outputs whether trained with conventional augmentations or AugUndo. Nonetheless, we still observe consistent (albeit smaller) improvements when FusionNet is trained with AugUndo.

For monocular depth estimation, we similarly evaluate Monodepth2, HR-Depth, and Lite-Mono models (trained on KITTI using standard augmentation pipelines and AugUndo) on Make3d. Tab. 5 shows that models trained on KITTI with AugUndo generalizes well to Make3d, gaining an average of around 8%

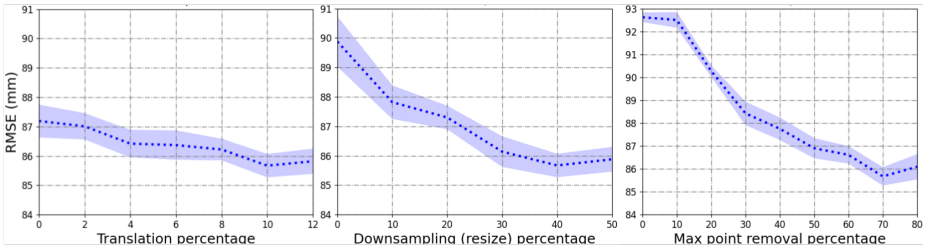

Fig. 2: *Effect of different degree of augmentation on VOID1500.*

improvement over all metrics and all models. Note: training Monodepth2 from their code repository reproduces their results on KITTI, but produces worse generalization results on Make3d than the weights they released; hence, we take the original weights released by the authors for evaluation. Nonetheless, we still improve over their best result.

J Sensitivity to Choice of Hyperparameters

In the main paper and Sec. A above, we described in details of the set of augmentations and degrees of each augmentation used in our experiments to achieve the reported numbers. We note that the performance gains can be obtained by typical set and ranges of augmentations and does not require a meticulous selection of hyper-parameters. In the paper, we tried to push the limits and tested increasing degrees of augmentation until performance saturated. Fig. 2 shows the trends of improvement for resizing (down), translation, and points removal. Choosing even small degrees of augmentations yields some performance gain. We note that there are diminishing returns as we scale up augmentations to large amounts, i.e. 50% or more in downsampling, which reduces the size of the observed image and predictions will have little to no detail. When computing the loss on these predictions on the original data, there is a considerable amount of noise and lends little to learning. Hence, at extreme degrees, we observe that performance eventually saturates. We also note that the choice of augmentation would depends on the task and dataset (i.e. vertical flip not applicable for KITTI since the camera is typical right-side-up for driving scenarios).

K Additional Results for Depth Completion

In the main paper, we demonstrated AugUndo for three unsupervised methods [43, 45, 48] on the VOID dataset. Due to space constraints, here, we show additional results on VOID1500 and VOID500 for a recent unsupervised distillation method, MonDi [22] as well as a recent unsupervised depth completion method, DesNet [50], in Tab. 6. Additionally, we show MonDi and DesNet for

VOID150 in Tab. 3. We applied photometric transformations of random brightness from 0.5 to 1.5, contrast from 0.5 to 1.5, saturation from 0.5 to 1.5, and hue from -0.1 to 0.1. We applied image patch removal by selecting between 0.1% to 0.5% of the pixels and removing 5×5 patches centered on them – approximately removing between 2.5% to 12.5% of the image. We applied random sparse depth point removal at a rate between 60% and 95% of all sparse points. We further applied geometric transformations of random horizontal and vertical flips, up to 10% of translation, between -25 to 25 degrees of rotation, and random resize factor between 0.6 to 1.1.

Table 6: *Quantitative results of MonDi, a distillation-based unsupervised method, on VOID.* Reported scores are mean and standard deviation over four independent trials. We evaluate MonDi trained on VOID1500, with standard augmentations and AugUndo, by testing them on VOID1500 and VOID500. Sparse depth maps within VOID1500 contains approximately 1500 points, and those within VOID500 contain approximately 3× less, respectively. AugUndo improves performance by an average of 5.33% and 9.02% across all methods and evaluation metrics on VOID1500 and VOID500, respectively.

	VOID1500				VOID500			
Method	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓
MonDi*	30.56±0.23	86.67±0.55	15.08±0.16	37.58±0.44	73.90±0.61	191.69±2.64	31.78±1.52	72.67±2.17
+ AugUndo	**29.02**±0.56	**79.34**±1.17	**14.56**±0.44	**35.93**±1.12	**69.90**±3.05	**157.02**±7.46	**29.61**±0.12	**68.48**±0.57
DesNet	37.41±0.28	93.31±0.76	19.17±0.24	45.57±0.62	74.89±0.67	170.32±2.03	35.62±0.72	78.30±1.31
+ AugUndo	**33.86**±0.60	**86.05**±0.43	**16.92**±0.29	**41.25**±0.31	**67.78**±1.10	**153.46**±2.64	**32.09**±0.48	**71.96**±1.05

Results on MonDi and DesNet. Overall, we improve MonDi by an average of 8.2% across all evaluation metrics across all sparsity levels (VOID1500, VOID500, and VOID150). This is surprising as amongst all the tested methods, MonDi is the most light-weight, with only 5.3M parameters. As we reduce model capacity, one would expect that the network to saturate in the data variations that can be modeled. However, despite the size of network is much smaller (23.2% less than KBNet), there is still a considerable gain when using AugUndo instead of their augmentation pipeline. This demonstrates efficacy and applicability of AugUndo; it can be used to improve methods with a range of capacities from tens of millions to several million. Also, we note that MonDi is an unsupervised distillation method (where it distills from unsupervised methods), so it is more closely related to supervised methods in supervision than unsupervised method. Even so, we observe a non-trivial improvement, which validates our discussion in Sec. 5 of the main paper regarding the applicability of AugUndo to *supervised* methods. Meanhwhile, we improves DesNet by an average of 10.64% accross all evaluation metrics, highlighting again AugUndo's applicability to future unsupervised depth completion methods.

Modeling AugUndo as change in camera pose versus camera parameters. In the main text, we discussed two ways of modeling our augmentation scheme: treating the augmented data as a result of changes in camera pose (i.e. motion) or

Table 7: *Quantitative results of KBNet with AugUndo as changes in camera pose or intrinsics (i.e. with and without depth adjustment, respectively) on VOID1500.* With depth adjustments (change in camera pose), the performance of the model is slightly worse (≈3.3% in terms of percent gain) than without depth adjustments (change in camera intrinsics), yet still better than the baseline by a large margin.

Method	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓	Gain (%)
KBNet	38.11±0.77	95.22±1.72	19.51±0.14	46.70±0.48	-
+ AugUndo (intrinsics)	**33.32**±0.18	**85.67**±0.39	**16.61**±0.29	**41.24**±0.60	**+12.29**
+ AugUndo (pose)	34.24±0.25	87.11±0.56	17.59±0.23	43.22±0.39	+8.99

in camera parameters (i.e. intrinsics). (1) Modeling the augmentation as camera motion requires adjusting sparse depth maps; for example, resizing can be treated as forward motion, so distance from camera to world surfaces need to be adjusted in the sparse depth map. Naturally, if choosing (1), then one would need to reproject the sparse depth points according to the camera motion. On the other hand, (2) modeling the augmentation as changes in the camera parameters does not require adjusting the sparse depth maps as the camera and 3D scene are both static; for example, resizing can now be treated as "zooming in" or an increase in focal length.

Tab. 7 compares the two approaches. We test the resizing operation by adjusting sparse depth, and likewise, dense output depth. Specifically, we assume a pin-hole camera model. As perspective projection is a linear, we adjust the depth value by scaling them using the random scale factor recorded during the random resizing operation. We observe the following: Firstly, both methods of modeling improves KBNet on VOID: modeling as (1) improves KBNet by 8.99% and (2) by 12.29%, respectively. This verifies the efficacy of AugUndo under both modeling choice. Secondly, modeling AugUndo as changes in (2) camera parameters improves over (1) camera motion. Particularly, (2) yields an additional 3.3% gain in the baseline and 3.6% relative improvement over KBNet trained using (1). This justifies our choice of modeling AugUndo as changes in camera parameters.

Naive geometric augmentations. In the main paper, we discussed that naively applying geometric augmentations can be detrimental to model performance and even prevent one from training them. As unsupervised depth completion and unsupervised monocular depth estimation assume rigid motion within the image triplet comprising of a training example, geometric augmentations that introduce some form of border padding (e.g., translation, rotation) will yield constant or edge extended borders across images (i.e., no motion in those regions). The lack of motion will result in PoseNet predicting near identity pose. Hence, naively incorporating geometric augmentations will prevent the model from properly learning depth and pose. This is demonstrated in Tab. 8 and 9 for rows marked with "+ Naive Geo Aug". Nonetheless, one can make modifications (no translation, center cropping on rotation, and resizing to the same shape for a batch) to ensure no borders are introduced during augmentation. Yet, we do not per-

Table 8: *Naive geometric augmentations for unsupervised depth completion on VOID.* It is infeasible to conduct unsupervised training with naive geometric augmentations. Modifying geometric augmentations to ensure no borders are introduced allows the model to train, but performance degrades. "+ Naive Geo Aug" denotes models trained with naive geometric augmentations and "+ Modified Geo Aug" denotes mdoels trained with modified geometric augmentations.

Method	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓
VOICED	74.78±2.69	139.75±4.57	39.20±1.46	71.98±2.54
+ Naive Geo Aug	554.65±91.10	642.13±69.40	639.42±113.11	980.81±78.01
+ Modified Geo Aug	109.40±13.01	205.14±18.22	106.21±13.27	286.71±31.79
+ AugUndo	**52.73**±0.41	**111.09**±0.92	**26.93**±0.54	**54.46**±0.38
FusionNet	52.11±0.44	113.30±1.18	28.53±0.52	58.79±2.01
+ Naive Geo Aug	82.73±6.96	190.71±9.90	73.61±33.32	224.62±156.04
+ Modified Geo Aug	57.72±6.67	138.52±15.87	31.91±5.19	74.41±14.64
+ AugUndo	**41.16**±0.18	**99.21**±0.39	**22.23**±0.35	**53.07**±1.30
KBNet	38.11±0.77	95.22±1.72	19.51±0.14	46.70±0.48
+ Naive Geo Aug	1,065.4 ± 81.7	1,216.4 ± 54.4	5,693.5 ± 401.3	7,032.7 ± 283.0
+ Modified Geo Aug	51.77 ± 4.13	118.95 ± 7.67	27.66 ± 3.04	60.92 ± 5.62
+ AugUndo	**33.32**±0.18	**85.67**±0.39	**16.61**±0.29	**41.24**±0.60

Table 9: *Naive geometric augmentations for unsupervised monocular depth estimation on KITTI.* It is infeasible to conduct unsupervised training when naively applying geometric augmentations. Modifying geometric augmentations to ensure no borders are introduced allows the model to train, but performance degrades. "+ Naive Geo Aug" denotes models trained with naive geometric augmentations and "+ Modified Geo Aug" denotes mdoels trained with modified geometric augmentations.

Method	RMSE ↓	Abs Rel ↓	Sq Rel ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Monodepth2	4.794 ± 0.035	0.117±0.001	0.845± 0.030	0.869± 0.004	0.959± 0.001	0.982±0.001
+ Naive Geo Aug	6.612±0.18	0.174±0.009	2.057±0.32	0.794±0.02	0.932±0.004	0.971±0.002
+ Modified Geo Aug	4.911±0.046	0.119±0.002	0.926±0.023	0.870±0.004	0.958±0.001	0.981±0.001
+ AugUndo	**4.739**±0.032	**0.113**±0.000	**0.862**±0.030	**0.879**±0.002	**0.960**±0.001	**0.982**±0.001
HR-Depth	4.626 ± 0.032	0.113±0.001	0.797± 0.022	0.879±0.002	0.961± 0.001	0.982±0.000
+ Naive Geo Aug	6.06±0.022	0.159±0.003	1.40±0.016	0.795±0.010	0.939±0.003	0.979±0.002
+ Modified Geo Aug	4.718±0.018	0.116±0.001	0.844±0.005	0.875±0.001	0.960±0.001	0.982±0.001
+ AugUndo	**4.610**±0.029	**0.111**±0.001	**0.794**±0.021	**0.883**±0.001	**0.962**±0.001	**0.983**±0.001

form the proposed "undo-ing" process. While this will allow the model to train, however, the effects of artifacts, loss during resampling, color distortion, and intensity saturation will lead to performance degradations worse than the baseline. This is shown in Tab. 8 and 9 for rows marked with "+ Modified Geo Aug".

L Additional Results on Monocular Depth Estimation

Applying our methods yields qualitative improvements in the depth prediction. In Fig. 3, we can see that our method better captures the building in the last column. Similarly, we observe similar improvements in the middle column for

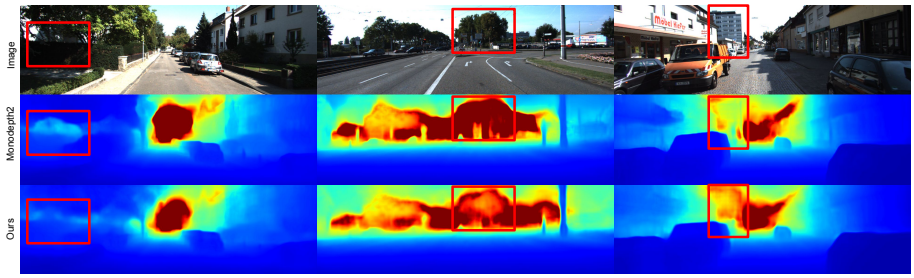

Fig. 3: Qualitative result of MonoDepth2 on KITTI. Red bounding boxes highlight areas where training with our augmentation scheme improves Monodepth2, i.e., wall and vegetation on left, trees in middle and building on right.

the trees that are located at a distance. This is in part thanks to augmentations such as resizing so that we can simulate objects at far and close distances. Additionally, training with AugUndo also improves over the baseline method on ambiguous regions such as the wall and vegetation in the first column, despite using the same hyper-parameter during training with the exception of data augmentation.

M Limitations

While we have proposed an algorithm to scale up augmentations for depth completion, a multimodal 3D reconstruction problem, admittedly our augmentations are limited to 2D. In the case of this inverse problem, the nuisance variability simulated by AugUndo are, in fact, projections of those in the 3D scene, which we do not model directly. Additionally, augmentations are used in other vision tasks including multiview stereo, binocular stereo, optical flow, etc. While we have shown that AugUndo can be applied to unsupervised and distillation (with dense supervision from pseudo ground truth), we have limited the scope of our method for depth completion, which considers a single image and sparse depth map as input. We foresee that AugUndo can also incorporate other modalities, such as tactile [51], be extended to other settings [24], and tasks, including deep feature visualization [19], semantic segmentation [3, 12, 20, 21, 32, 41], and object detection [2, 15, 17, 28, 35]. Additionally, we hypothesize that AugUndo can also be extended towards multi-frame geometric tasks such as stereo [1, 40, 46, 49], optical flow [16, 18, 33, 36, 56, 57], etc., but one must account for their specifics and problem setups, i.e. stereo assumes frontoparallel views. Lastly, like all scaling problems, AugUndo is eventually limited by diminishing returns. As certain augmentation are pushed to extremes, i.e., maximum brightness such that the image is "white", large spatial reduction such that the image is small, there is little to no information in the input to infer depth; even if the supervision signal exists, it would be mapping nonsensical inputs to depth maps, which does

not improve performance (as illustrated by Fig. 2), and lending to saturation in performance gain.

References

1. Berger, Z., Agrawal, P., Liu, T.Y., Soatto, S., Wong, A.: Stereoscopic universal perturbations across different architectures and datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2022)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
5. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
6. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
7. Ezhov, V., Park, H., Zhang, Z., Upadhyay, R., Zhang, H., Chandrappa, C.C., Kadambi, A., Ba, Y., Dorsey, J., Wong, A.: All-day depth completion. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2024)
8. Fei, X., Wong, A., Soatto, S.: Geo-supervised visual depth prediction. IEEE Robotics and Automation Letters **4**(2), 1661–1668 (2019)
9. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
10. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
11. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
12. Han, M., Bushong, E.A., Segawa, M., Tiard, A., Wong, A., Brady, M.R., Momcilovic, M., Wolf, D.M., Zhang, R., Petcherski, A., et al.: Spatial mapping of mitochondrial networks and bioenergetics in lung cancer. Nature **615**(7953), 712–719 (2023)
13. Harris, C.G., Stephens, M., et al.: A combined corner and edge detector. In: Alvey vision conference. vol. 15, pp. 10–5244. Citeseer (1988)
14. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic gradient descent. In: ICLR: International Conference on Learning Representations. pp. 1–15. ICLR US. (2015)
15. Lao, D., Sundaramoorthi, G.: Minimum delay moving object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4250–4259 (2017)

16. Lao, D., Sundaramoorthi, G.: Extending layered models to 3d motion. In: Proceedings of the European conference on computer vision (ECCV). pp. 435–451 (2018)
17. Lao, D., Sundaramoorthi, G.: Minimum delay object detection from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5097–5106 (2019)
18. Lao, D., Wang, C., Wong, A., Soatto, S.: Diffeomorphic template registration for atmospheric turbulence mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25107–25116 (2024)
19. Lao, D., Wu, Y., Liu, T.Y., Wong, A., Soatto, S.: Sub-token vit embedding via stochastic resonance transformers. In: International Conference on Machine Learning. PMLR (2024)
20. Lao, D., Yang, F., Wang, D., Park, H., Lu, S., Wong, A., Soatto, S.: On the viability of monocular depth pre-training for semantic segmentation. In: European Conference on Computer Vision. Springer (2024)
21. Liu, C., Chen, L.C., Schroff, F., Adam, H., Hua, W., Yuille, A.L., Fei-Fei, L.: Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 82–92 (2019)
22. Liu, T.Y., Agrawal, P., Chen, A., Hong, B.W., Wong, A.: Monitored distillation for positive congruent depth completion. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II. pp. 35–53. Springer (2022)
23. Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y.: Hr-depth: High resolution self-supervised monocular depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2294–2301 (2021)
24. Park, H., Gupta, A., Wong, A.: Test-time adaptation for depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20519–20529 (2024)
25. 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)
26. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
27. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
28. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
29. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence **31**(5), 824–840 (2008)
30. 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)
31. Singh, A.D., Ba, Y., Sarker, A., Zhang, H., Kadambi, A., Soatto, S., Srivastava, M., Wong, A.: Depth estimation from camera image and mmwave radar point cloud.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9275–9285 (2023)
32. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)
33. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8934–8943 (2018)
34. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
35. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
36. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
37. 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)
38. Upadhyay, R., Zhang, H., Ba, Y., Yang, E., Gella, B., Jiang, S., Wong, A., Kadambi, A.: Enhancing diffusion models with 3d perspective geometry constraints. ACM Transactions on Graphics (TOG) **42**(6), 1–15 (2023)
39. 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)
40. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: Learned multi-view patchmatch stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14194–14203 (2021)
41. Wong, A., Chen, A., Wu, Y., Cicek, S., Tiard, A., Hong, B.W., Soatto, S.: Small lesion segmentation in brain mris with subpixel embedding. In: International MICCAI Brainlesion Workshop. pp. 75–87. Springer (2021)
42. Wong, A., Cicek, S., Soatto, S.: Targeted adversarial perturbations for monocular depth prediction. Advances in neural information processing systems **33**, 8486–8497 (2020)
43. 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)
44. 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)
45. Wong, A., Fei, X., Tsuei, S., Soatto, S.: Unsupervised depth completion from visual inertial odometry. IEEE Robotics and Automation Letters **5**(2), 1899–1906 (2020)
46. Wong, A., Mundhra, M., Soatto, S.: Stereopagnosia: Fooling stereo networks with adversarial perturbations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2879–2888 (2021)
47. 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)

48. 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)
49. Xu, H., Zhang, J.: Aanet: Adaptive aggregation network for efficient stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1959–1968 (2020)
50. Yan, Z., Wang, K., Li, X., Zhang, Z., Li, J., Yang, J.: Desnet: Decomposed scale-consistent network for unsupervised depth completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3109–3117 (2023)
51. Yang, F., Feng, C., Chen, Z., Park, H., Wang, D., Dou, Y., Zeng, Z., Chen, X., Gangopadhyay, R., Owens, A., et al.: Binding touch to everything: Learning unified multimodal tactile representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26340–26353 (2024)
52. 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)
53. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5684–5693 (2019)
54. Zeng, Z., Wang, D., Yang, F., Park, H., Soatto, S., Lao, D., Wong, A.: Wordepth: Variational language prior for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9708–9719 (2024)
55. Zhang, N., Nex, F., Vosselman, G., Kerle, N.: Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18537–18546 (2023)
56. Zhang, X., Pak, D.H., Ahn, S.S., Li, X., You, C., Staib, L., Sinusas, A.J., Wong, A., Duncan, J.S.: Heteroscedastic uncertainty estimation for probabilistic unsupervised registration of noisy medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2024)
57. Zhang, X., Stendahl, J.C., Staib, L., Sinusas, A.J., Wong, A., Duncan, J.S.: An adaptive correspondence scoring framework for unsupervised image registration of medical images. In: European Conference on Computer Vision. Springer (2024)
58. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017)