AugUndo: Scaling Up Augmentations for Monocular Depth Completion and Estimation

Yangchao Wu[1], Tian Yu Liu[1], Hyoungseob Park[2], Stefano Soatto[1], Dong Lao[1], and Alex Wong[2]

[1] UCLA Vision Lab, Los Angeles, CA 90095, USA
{wuyangchao1997, tianyu, soatto, lao}@cs.ucla.edu
[2] Yale Vision Lab, New Haven, CT 06511, USA
{hyoungseob.park, alex.wong}@yale.edu

Abstract. Unsupervised depth completion and estimation methods are trained by minimizing reconstruction error. Block artifacts from resampling, intensity saturation, and occlusions are amongst the many undesirable by-products of common data augmentation schemes that affect image reconstruction quality, and thus the training signal. Hence, typical augmentations on images viewed as essential to training pipelines in other vision tasks have seen limited use beyond small image intensity changes and flipping. The sparse depth modality in depth completion have seen even less use as intensity transformations alter the scale of the 3D scene, and geometric transformations may decimate the sparse points during resampling. We propose a method that unlocks a wide range of previously-infeasible geometric augmentations for unsupervised depth completion and estimation. This is achieved by reversing, or "undo"-ing, geometric transformations to the coordinates of the output depth, warping the depth map back to the original reference frame. This enables computing the reconstruction losses using the original images and sparse depth maps, eliminating the pitfalls of naive loss computation on the augmented inputs and allowing us to scale up augmentations to boost performance. We demonstrate our method on indoor (VOID) and outdoor (KITTI) datasets, where we consistently improve upon recent methods across both datasets as well as generalization to four other datasets. Code available at: https://github.com/alexklwong/augundo

Keywords: depth completion · monocular depth · augmentations

1 Introduction

Data augmentation is essential to training machine learning models; it plays a role in performance and generalization [40, 51, 55]. One common axiom of choosing augmentations is that the task output should remain invariant to the augmentation. For example, image flipping is a viable augmentation for classifying animals, since it does not alter the label. Conversely, flipping road signs can alter their meanings; hence, such augmentation can be detrimental to tasks involving road sign recognition. For geometric tasks, the range of augmentations is more restricted due to constraints of problem formulation: Stereo assumes pairs of frontoparallel rectified images; hence, in-plane rotations are not viable. Image-guided sparse depth completion relies on sparse points to ground estimates to metric scale; therefore, intensity transformations on sparse depth maps that alter the scale of the 3-dimensional (3D) scene are infeasible. Unsupervised learning of depth completion and estimation further limit the use of augmentations as the supervision signal comes from reconstructing the inputs, where augmenting the input introduces artifacts that impact reconstruction quality and therefore the supervision (see Sec. D, Fig. 1, Tab. 8, 9 in the Supp. Mat. for examples of artifacts and extended discussion on their effect on learning). Moreover, video-based unsupervised training assumes rigid motion, so augmentations that introduce padding (e.g., translation, rotation) will yield constant or edge extended borders across images (i.e., no motion), preventing the model from properly learning depth and pose – leaving few augmentations viable. While simulating nuisances is desirable, naively applying augmentations may do more harm than good; thus, it is unsurprising that existing work in unsupervised depth completion [33, 36, 50, 63–65, 67, 73] and estimation [14, 15, 35, 81, 85] primarily rely on a small range of photometric augmentations and flipping.

Nevertheless, photometric augmentations help model the diverse range of illumination conditions and colors of objects that may populate the scene; geometric augmentations can simulate the various camera parameters, i.e., image resizing (zooming) can model changes in focal length, and scene arrangements, i.e., image flipping. However, block artifacts, loss during resampling, and intensity saturation are just some of the many undesirable side-effects of traditional augmentations to the image and sparse depth map for unsupervised learning of geometric tasks. To avoid compromising the supervision signal, we compute the typical reconstruction loss on the original input image and sparse depth map instead of the augmented inputs, which bypasses negative effects of reconstruction artifacts due to photometric and geometric augmentations. However, there exists a mis-alignment between the original input (e.g., image, sparse depth), and the model depth estimate as geometric augmentations induce a change in coordinates. Hence, we *undo* the geometric augmentations by inverting them in the output space to align the model estimate with the training target.

Amongst the many geometric tasks, we focus on *unsupervised depth completion*, the task of inferring a dense depth map from an image and sparse depth map, where augmentations have seen limited use. Here, a training sample includes the input sparse depth map, its associated image and additional images of neighboring views of the same 3D scene. Our method is also applicable to *unsupervised monocular depth estimation*, which omits the sparse depth modality. Augmentations have traditionally been restricted to a limited range of photometric transformations and flipping – due to the need to preserve photometric consistency across a sequence of video frames used during training, and the sparse set of 3D points projected onto the image frame as a 2.5D range map; degradation to either modalities directly impacts the supervision signal as the loss function is conventionally computed on the augmented inputs. By using our

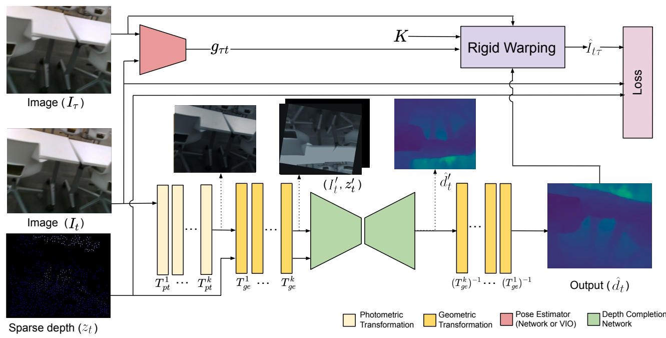

Fig. 1: *Overview.* We apply photometric augmentations to the input image, and the same set of geometric augmentations to both the input image and sparse depth map. We warp the output depth back to the original reference frame with the inverse geometric transformations. This enables image and sparse depth reconstruction losses to be computed on the original inputs while unlocking previously-infeasible augmentations.

method, loss functions involving sparse depth and image reconstruction from other views can be computed on the original inputs while applying augmentations that were previously not viable for the task. **Our hypothesis:** By "undoing" the augmentations, one can expand the viable set and scale up their use in training, leading to improved model performance and generalizability.

To this end, we introduce AugUndo, an augmentation framework that enables one to apply a wide range of photometric and geometric transformations on the inputs, and to "undo" them during loss computation. This allows computing the unsupervised loss on the original images and sparse depth maps, free of artifacts, through a warping of the output depth map – obtained from augmented input – onto the input frame of reference based on the inverse geometric transformation. In addition to group transformations that allow for output alignment, we combine them with commonly employed photometric augmentations. To the best of our knowledge, we are the first to propose a unified augmentation scheme for photometric and geometric augmentations for unsupervised depth completion and estimation. We demonstrate AugUndo on recent methods on indoor and outdoor settings, where we consistently improve all methods across all datasets.

Our contributions are as follows: (1) We propose AugUndo, a simple-yet-effective framework to scale up photometric and geometric augmentations for unsupervised depth completion and estimation, without compromising the supervision signal; AugUndo can be applied in a plug-and-play manner to existing methods with negligible increase in computational costs during training. (2) We enable previously-infeasible augmentations to be used for training unsupervised methods and comprehensively ablate combinations of eleven types of augmentations to study the performance benefits of each. (3) We show that AugUndo can consistently improve robustness to shifts in sparse point densities for completion, model performance as well as zero-shot generalization for both depth completion and estimation for indoor and outdoor scenarios; thus, validating our hypothesis.

2 Related Work

Data augmentation for depth completion and estimation. While photometric transformations such as color jitter are often applied for unsupervised depth completion [36,63–65], and estimation [10,14–16,26,27,41,47,66,79,80], geometric augmentations other than simple flipping are less commonly adopted. Most supervised depth completion methods [8,18,19,21,23,29,31,43–45,69,73,82,83] similarly limit their augmentations to color jitter and horizontal flips due to sparse depth maps being decimated by rotation and scaling, causing points to be interpolated away. Nevertheless, for supervised training, it is straightforward to directly apply the same transformation to the ground truth for training. Indeed, some supervised depth completion methods [2,3,9,38,39,53,59,78] adopted random scaling, translation, in-plane rotation. Translation augmentations are also commonly applied in supervised depth estimation methods [6, 7, 25, 30, 32, 46, 47, 58, 62, 68, 71, 76]. We posit that the reason that such transformations can be applied to the ground truth in supervised settings is due to artifacts caused by the transformation of a piece-wise smooth depth map being less severe than those of an RGB image and its intensities. However, such assumptions do not hold in unsupervised training. These artifacts would affect the training signal for unsupervised methods, which rely on photometric correspondences and observed sparse points. Our approach aims to resolve this to allow diverse geometric augmentations to be applied in a plug-and-play fashion in unsupervised training.

Unsupervised depth completion methods [36, 50, 63–65, 70, 73] leverage image and sparse depth reconstructions as training signals by minimizing errors between the input image and its reconstruction from other views, and errors between the input sparse depth map and the predicted depth along with a local smoothness regularizer. [36] used Perspective-n-Point [28] and RANSAC [11] to align consecutive video frames. [73] learns a depth prior conditioned on the image. [63] uses synthetic data to learn a prior on the shapes populating a scene, while [34] translates synthetic data to real domain to leverage rendered depth. [64] proposed an adaptive scheme to reduce penalties incurred on occluded regions. [67] maps the image onto the 3D scene through calibrated backprojection. [70] decouples structure and scale. [20] uses line features from visual SLAM and [33] introduced monitored distillation for positive congruent training. Augmentations for these methods are limited to a small range of photometric perturbations and image flipping. Operations such as rotation, resizing, and translation require resampling, which creates artifacts and affects the reconstruction quality. This causes performance degradation since loss is typically computed on the augmented images. Loss in sparse depth maps is further impacted as resampling and interpolation may cause loss of sparse points. Contrary to these limited augmentation schemes, we enable a large range of photometric and geometric augmentations to be introduced during training.

Unsupervised monocular depth estimation, like depth completion, also minimizes photometric reconstruction error. [12] frames depth estimation as a novel view synthesis problem. [14] improves [12] by imposing a consistency loss on the depth predicted from left and right images. [85] uses a pose network to enable unsupervised training on video sequences. [15] introduced auto-masking and min reprojection loss. To improve the supervision signal, [4, 56, 61] leverage noisy proxy labels and [42] uses a trinocular assumption. Additional loss terms based on visual odometry [60], iterative closest point [37], surface normals [75], and semantic segmentation [17, 24] were also introduced; [41, 72] further included predictive uncertainty. To handle rigid and non-rigid motions in the scene, previous works explored multi-task learning to include optical flow and moving object estimation [1, 48, 74, 77, 86], and used semantics to filter out outlier regions [22]. [35] redesigned the skip connection and decoders to extract high-resolution features, [84] combined global and local representations and [81] introduced a lightweight architecture with dilated convolution and attention. In addition to depth completion, our method also shows consistent improvements on monocular depth estimation for [15, 35, 81].

3 Method Formulation

Let $I : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3_+$ be an RGB image captured by a calibrated camera, $z : \Omega_z \subset \Omega \rightarrow \mathbb{R}_+$ the corresponding sparse point cloud projected onto the image plane as a sparse depth map, and $K \in \mathbb{R}^{3\times 3}$ the intrinsic calibration matrix. Given an image and its sparse depth map, depth completion aims to learn a function $f_\theta(I, z)$ that recovers the distance between the camera to points in the 3D scene as a dense depth map. In another mode, if sparse depth maps are not given, then the problem reduces to monocular depth estimation which learns a function $f_\theta(I)$ to map a single image to a depth map $\Omega \rightarrow \mathbb{R}_+$. For the ease of notation, we denote the output depth map for both depth completion and estimation as $\hat{d} \in \mathbb{R}^{H\times W}_+$ where H and W are its height and width.

Unsupervised depth completion relies on photometric and sparse depth reconstruction errors as its primary supervision signal. Without loss of generality, we assume an input of (I_t, z_t) for an RGB image and associated sparse depth map captured at time t and during training, an additional set of temporally adjacent images I_τ for $\tau \in \Upsilon \doteq \{t-1, t+1\}$. The reconstruction $\hat{I}_{t\tau}$ of I_t from image I_τ is given by the reprojection based on estimated depth $\hat{d}_t := f_\theta(\cdot)$

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

where $\bar{x} = [x^\top, 1]^\top$ is the homogeneous coordinates of $x \in \Omega$, $g_{\tau t} \in SE(3)$ the relative pose of the camera from time t to τ, K the intrinsic calibration matrix, and π the canonical perspective projection.

Using Eq. (1), a depth completion or estimation network f_θ minimizes

$$\arg\min_\theta \sum_{\tau\in\Upsilon} \sum_{x\in\Omega} \alpha\rho\big(\hat{I}_{t\tau}(x), I_t(x)\big) + \sum_{x\in\Omega_z} \beta\psi\big(\hat{d}_t(x), z_t(x)\big) + \lambda R(\hat{d}_t) \tag{2}$$

where ρ denotes the photometric reconstruction error, typically L_1 difference in pixel values and/or structural similarity (SSIM), ψ the sparse depth reconstruction error, typically L_1 or L_2, R the regularizer that biases the depth map to be piece-wise smooth with depth discontinuities aligned with edges in the image (commonly used by previous works [15,36,65,67]), and α, β and λ their respective weighting. Note: monocular depth omits sparse depth term, i.e., $\beta = 0$.

Inverse transformation. Let $\mathcal{A}_{pt}$ be the set of possible photometric transformations, and $\mathcal{A}_{ge}$ be the set of all geometric transformations. Given $T_{ge} \in \mathcal{A}_{ge}$, we wish to obtain an inverse transformation T_{ge}^{-1} such that $T_{ge} \circ T_{ge}^{-1} \approx Id$ the identity function[3]. At each time step, we can sample a sequence of transformations $\{T_{pt}^1 \dots T_{pt}^k\}$ and $\{T_{ge}^1 \dots T_{ge}^m\}$ respectively from $\mathcal{A}_{pt}$ and $\mathcal{A}_{ge}$ to construct transformations $T_{pt} = T_{pt}^1 \circ \dots \circ T_{pt}^k$ and $T_{ge} = T_{ge}^1 \circ \dots \circ T_{ge}^m$. We denote the composition of a collection of augmentation transformations by $T = T_{ge} \circ T_{pt}$ where T_{pt} denotes photometric transformation, and T_{ge} denotes geometric transformation. Furthermore, we denote the inverse geometric transformations by $T_{ge}^{-1} = (T_{ge}^m)^{-1} \circ (T_{ge}^{m-1})^{-1} \circ \dots \circ (T_{ge}^1)^{-1}$, which operates on the space of depth maps to reverse the geometric transformation so that we can warp the output depth map onto the reference frame of the original image. Through inverse warping, this process is differentiable. In practice, some transformations cause border regions of the image to be warped out of frame, i.e., translated off the image plane or cropped away. Hence, after warping our output depth map back to the original frame of reference using the inverse geometric transformations, border extensions (edge paddings) are introduced to handle out-of-frame regions.

AugUndo. We apply each geometric transformation over image coordinates and resample (bilinear for image and nearest neighbor for sparse depth):

$$\begin{bmatrix} x' & 1 \end{bmatrix}^\top = T_{ge} \begin{bmatrix} x & 1 \end{bmatrix}^\top \tag{3}$$

$$I'(x') = T_{pt}(I)(x); \;\; z'(x') = z(x) \tag{4}$$

where T_{ge} is the geometric transformation, $x \in \Omega$ and $x' \in \Omega$ are coordinates in the image grid, and I' is the image after the transformation; for ease of notation, we hereafter denote $I' = T(I) = T_{ge} \circ T_{pt}(I)$ to include both augmentations through composition. Note that x is in the original image reference frame and x' is in the transformed image reference frame. Naturally, this process can be extended to multiple geometric transformations by composing them, i.e., $T_{ge} = T_{ge}^1 \circ T_{ge}^2 \circ \dots \circ T_{ge}^m$. The reverse process is simply inverting the transformations where $T_{ge}^{-1} = (T_{ge}^m)^{-1} \circ (T_{ge}^{m-1})^{-1} \circ \dots \circ (T_{ge}^1)^{-1}$:

$$\begin{bmatrix} x'' & 1 \end{bmatrix}^\top = T_{ge}^{-1} \begin{bmatrix} x' & 1 \end{bmatrix}^\top \tag{5}$$

$$\hat{d}(x'') = \hat{d}'(x') \tag{6}$$

where $\hat{d}'$ is the depth map inferred from augmented inputs (I', z'). Once reverted back to the original reference frame, $\hat{d}$ is used to reconstruct I_t from I_τ for $\tau \in \Upsilon$ as in Eq. (1). More details can be found in Alg. 1 in the Supp. Mat.

By modeling T_{ge}^{-1}, Eq. (3)-Eq. (6) allow us to apply a wide range of augmentations, while still establishing the correspondence between I_t and I_τ. Specifically,

[3] Note that in practice, not all transformations are bijective. For instance, image rescaling with a fixed size image plane; hence strict equality is not always possible.

for minimizing Eq. (2), one may simply augment the image and sparse depth by T to yield (I_t', z_t') as input, and reconstruct the original image and sparse depth (I_t, z_t) from other views I_τ and the aligned output depth $\hat{d}$ (Eq. (6)). We note that the inverse transformation is critical for enabling the sparse depth reconstruction term to be computed properly in Eq. (2); if computed in the transformed reference frame, i.e., on z_t', many of the sparse points would be decimated by interpolation during rotation and resizing (downsampling) augmentations – leaving a lack of supervision from the sparse depth term of the loss function. We note that the original image sequence is fed to the pose estimator, rather than the augmented images. This ensures the estimated relative pose used during reprojection (Eq. (1)) also corresponds to the original images.

Modeling geometric augmentations. Depth completion differs from typical image-based (e.g., monocular) depth estimation, where operations like resizing are susceptible to scale ambiguity and can either correspond to changes in focal length, or distance from the camera (i.e., due to changes in camera pose). In depth completion, the sparse depth map grounds image pixels to specific depth values. Hence, there is no ambiguity in distance of each sparse point from the camera. Our goal is to synthesize new training data via geometric transformations that are consistent with the original 3D scene. Given that there is no scale ambiguity, this can be achieved by either (1) changes in camera pose (extrinsics), or (2) changes in camera parameters (intrinsics).

Augmentation via (1) would require warping the inputs for view synthesis. Since we operate under the unsupervised setting, accurate warping (preserving 3D scene) without access to dense ground truth depth values is difficult to achieve. An exception to this is rotation, which we address below. On the other hand, (2) can be achieved via standard 2D image transformations while preserving the 3D structure of the scene. This motivates our modeling of geometric transformations, i.e., resizing, translation, as changes in camera parameters.

For example, augmenting focal length (i.e. zooming in/out) is akin to "resizing". Similarly, translation can be used to model shifting of the optic center, i.e. capturing the same 3D scene using cameras with different principal point offsets. Note: image rotation can be seen as rotating a camera about its optical axis, which does not change the distance of a point from the camera. Thus, the camera position and the scene are kept constant across augmentations, eliminating the need for adjusting depth values via warping during the process. Instead, we only have to realign the output depth back to the original frame of reference (Eqs. (5) and (6)), which makes our method computationally efficient.

Augmentations. Our choices are based on common nuisance variabilities, i.e., changes in illumination, occlusions, and object color, scene layout (flip), and camera parameters (resize, translation) and orientation (rotation).

Photometric. We include brightness, contrast, saturation, and hue, where all follow standard augmentation pipeline in existing works [15,65,67,81]. Applying the inverse of photometric augmentation can be viewed as recovering the original image; hence, we directly use original image instead of applying transformations to the intensities (which are not recoverable if saturated at the max value).

Occlusion. We consider patch removal and sparse point removal. For the former, we randomly select a percentage of pixels $x \in \Omega$ in the image and remove (with zero-fill) an arbitrary-sized patch around it. For the latter, we randomly sample a percentage of points $x \in \Omega_z$ in the sparse depth map. The inverse transformation of this is simply the original image and sparse depth map, both used in loss computation before augmentation.

Flip. We consider horizontal and vertical flips. When applied, the same flip operation is used for both input image and sparse depth maps. We record the flip type during data augmentation. During loss computation, we reverse the flip direction to align output depth with the original image and sparse depth map.

Resize. We define a new image plane of the input resolution and generate a random scaling factor to be applied along both height and width directions. The image is warped to the new image plane, where any point warped out of the plane is excluded; borders of the warped image are extended to the bounds of the image plane by edge replication. Sparse points occupying multiple pixels are eroded to a single point for resizing with a factor greater than 1. During loss computation, we warp the output depth map onto a new image plane of the same dimensions by the inverse scaling matrix of the recorded scaling factors; borders of the warped depth map are extended by edge replication if necessary. We view resizing factors greater and less than 1 (zooming in and out) as two separate forms of augmentations to distinguish their contributions.

Rotation. Naive rotation leads to the loss of large areas of the image, i.e., cropped away to retain the same-size image, discarding large portions of possible co-visible regions and also supervisory signals. To preserve the entire image, we first warp the image by a randomly generated angle to a new (larger) image plane, so that the rotated image fits tightly within the new image. As image sizes within a batch can vary depending on the rotation angle, we center-pad (uniformly on each side) each image in the batch to the maximum width and height of the augmented batch. To reverse the rotation on the output depth, we warp the output depth map back by the inverse rotation matrix, then perform a center crop on the depth map to align with the original input.

Translation. We define a new image plane of the same dimensions as the input and generate random height and width translation factors. The coordinates of the input are translated and its pixels are inverse warped onto the new image plane. Any pixel warped out of the image plane is excluded. Borders of the warped image are extended to the bounds of the image plane by edge replication. During loss computation, we warp the output depth map onto a new image plane of the same dimensions by the inverse translation matrix. Borders of the warped depth map are extended to the bounds of the image plane by edge replication.

4 Experiments

We demonstrate AugUndo on recent unsupervised depth completion (VOICED [65], FusionNet [63], KBNet [67]) and estimation (Monodepth2 [15], HRDepth [35], LiteMono [81]) methods on two datasets (KITTI [13, 57], VOID [65]) here.

Due to space limitations, we defer the results of MonDi [33] and DesNet [70] to Sec. K in the Supp. Mat. To evaluate zero-shot generalization, we test on three additional datasets for each task: NYUv2 [52], ScanNet [5], Waymo [54], and Make3d [49] – where we transfer models on VOID to NYUv2 (Tab. 3) and ScanNet, and from KITTI to Waymo (Tab. 4 in Supp. Mat.) for completion and Make3d [49] (Tab. 5 in Supp. Mat.) for estimation. We also test model sensitivity to different sparse depth input densities in Tab. 1 (right) – see Sec. H of the Supp. Mat for an extensive study. Additionally, we present an comprehensive ablation study in Sec. G in the Supp. Mat. to test the effect of each augmentation. We also provide results on modeling AugUndo as changes in camera motion or parameters in Tab. 7 of the Supp. Mat.; interpreting geometric augmentations as different camera parameters yields better performance. Tabs. 8, 9 in Supp. Mat. show that unsupervised models cannot be trained with naive geometric augmentations. See Sec. E of the Supp. Mat. for descriptions of datasets.

Experiment setup. We followed the original settings of the open-sourced code for each work and modified the data handling and loss function to incorporate AugUndo. We perform 4 independent trials for each experiment and report their means and standard deviations. To ensure a fair comparison, we train all models from scratch. Below, we report the best combination of augmentations found empirically through extensive experiments on each dataset. See Secs. A and B in Supp. Mat. for implementation details, and evaluation metrics, respectively.

Augmentations. Through a search over augmentation types and values, we found a consistent set that tends to yield improvements across all methods with small changes to degree of augmentation catered to each method. See Sec. C in Supp. Mat. for details of augmentation parameters. Rows in Tabs. 1 to 6 marked with "+ AugUndo" denote models trained with our method. 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 Sec. J of Supp. Mat. for a sensitivity study). Results reported here aims to study how far one can push performance and generalization, purely from augmentations.

Results on VOID. We begin by presenting quantitative results for AugUndo on unsupervised depth completion. Tab. 1 (left, VOID1500) shows our main results on the VOID benchmark. By training the models with AugUndo, we observe an average overall improvement of 18.3% across all methods and metrics on VOID1500. Specifically, we improve VOICED by 26.4%, FusionNet by 16.3%, and KBNet by 12.2%. These experiments validate our hypothesis that by applying a wider range of augmentations, we are able to improve the baseline model's performance. They also illustrate the lack in use of augmentations in existing works: incorporating standard augmentations (albeit requiring modification to augmentation and loss computation pipelines) can yield a large performance boost. Fig. 2 shows a head-to-head comparison between KBNet trained using standard procedure in [67] and KBNet trained using AugUndo. We observe qualitative improvements from AugUndo where we improve KBNet in homogeneous regions and image borders, i.e., pillar (left), cabinet (middle), wall (right). Applying geometric augmentations yields models with fewer border

Table 1: *Depth completion on VOID.* AugUndo improves performance by an average of 18.3% across all methods and metrics on VOID1500. When models trained on VOID1500 are tested VOID500, AugUndo improves by 23.1%, as translation, resizing, and occlusion augmentations vary sparsity by removing sparse points from the input.

	VOID1500				VOID500			
Method	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓
VOICED	74.78±2.69	139.75±4.57	39.20±1.46	71.98±2.54	137.01±4.23	235.80±7.82	71.36±1.86	130.63±5.66
+ AugUndo	**52.73**±0.41	**111.09**±0.92	**26.93**±0.54	**54.46**±0.38	**92.99**±1.11	**176.94**±1.38	**46.43**±0.85	**91.10**±1.64
FusionNet	52.11±0.44	113.30±1.18	28.53±0.52	58.79±2.01	97.73±0.73	194.32±1.36	58.65±1.31	122.95±3.04
+ AugUndo	**41.16**±0.18	**99.21**±0.39	**22.23**±0.35	**53.07**±1.30	**74.97**±0.69	**162.71**±2.39	**40.44**±1.39	**92.11**±5.79
KBNet	38.11±0.77	95.22±1.72	19.51±0.14	46.70±0.48	78.44±1.39	178.17±3.27	37.56±0.61	83.43±1.89
+ AugUndo	**33.32**±0.18	**85.67**±0.39	**16.61**±0.29	**41.24**±0.60	**66.97**±0.81	**151.55**±2.03	**31.63**±0.53	**71.90**±0.82

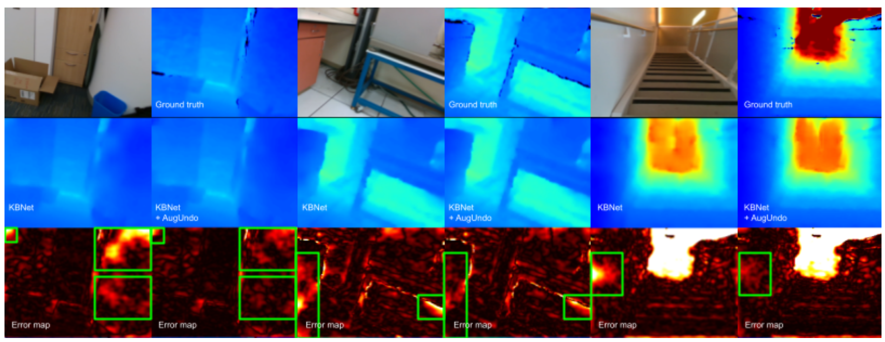

Fig. 2: *Depth completion on VOID.* We compare KBNet trained with standard augmentations and with AugUndo. AugUndo consistently produces lower errors with reduced border artifacts and improves on homogeneous regions, i.e., pillar (left), cabinet (middle), wall of staircase (right). Error maps are highlighted for comparison.

artifacts as we apply random translation, which warps part of the image out of frame. This allows us to simulate border occlusion, where there lacks correspondence in an adjacent frame. While training with standard protocol results in failures to recover structures near the image border, training with AugUndo can render models robust to them by computing the loss on the original frame of reference (where we do have correspondence) through our inverse transformations of the depth map. Additionally, resizing allows the model to learn multiple resolutions of the input, akin to zooming in and out, which can also impose smoothness in homogeneous regions through the scale-space transition; whereas, rotation can simulate diverse camera orientations. Fig. 2 shows that translation, rotation, and resizing can model these effect in the input space to yield models robust to these nuisance variability.

AugUndo also is applicable for monocular depth estimation. Tab. 2 shows a comparison using the standard augmentation procedure of Monodepth2, HR-Depth, and Lite-Mono and using AugUndo. Tab. 2 shows that AugUndo consistently improves all models across all error and accuracy metrics. Thus, validating

Table 2: *Monocular depth estimation on VOID.* AugUndo is applicable for monocular depth estimation and consistently improves three monocular depth models.

Method	MAE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2	283.861±3.732	395.947±5.728	0.183±0.002	0.100±0.003	0.717±0.005	0.922±0.004	0.975±0.002
+ AugUndo	**277.696**±4.861	**388.088**±5.768	**0.178**±0.003	**0.095**±0.004	**0.724**±0.007	**0.925**±0.004	**0.978**±0.002
HR-Depth	286.282±7.059	399.112±9.184	0.185±0.004	0.100±0.004	0.714±0.012	0.919±0.006	0.975±0.002
+ AugUndo	**283.086**±6.787	**394.261**±9.133	**0.181**±0.005	**0.097**±0.005	**0.718**±0.013	**0.922**±0.004	**0.977**±0.002
Lite-Mono	319.910±15.00	446.005±22.97	0.209±0.013	0.129±0.019	0.669±0.016	0.892±0.011	0.963±0.006
+ AugUndo	**308.010**±0.859	**426.626**±0.484	**0.200**±0.003	**0.114**±0.001	**0.674**±0.005	**0.901**±0.002	**0.969**±0.002

the hypothesis that AugUndo can be applied generically to improve monocular depth estimation. Specifically, we observe a boost in the most difficult accuracy metric ($\delta < 1.25$), where Monodepth2 improves from 0.717 to 0.724, HR-Depth from 0.714 to 0.718 and Lite-Mono from 0.669 to 0.674. At the same time, for AbsRel metric, Monodepth2 improves from 0.183 to 0.178, HR-depth improves from 0.185 to 0.181, and Lite-Mono improves from 0.209 to 0.200.

Sensitivity study. One limitation of existing depth completion training pipelines is that there are little to no augmentations applied to sparse depth modality. However, in real-world applications, sparse depth has, in fact, high variability, i.e., features tracked in SLAM/VIO systems will vary depending on the scene (points are dropped or added to the state), and point cloud densities returned by a sensor will differ based on specifications. To further examine the effect of AugUndo on sparse depth, we study the sensitivity to changes in sparse depth density by testing models on VOID1500 (≈1500 points) on VOID500 (≈500). For the 3× reduction in sparse points, AugUndo improves robustness by an average of 23.1% across all methods. These improvements result from the geometric and occlusion augmentations made possible via AugUndo, which greatly increases sparse depth variations, i.e., decimating them through resizing, reorienting their configuration through rotation, translating them out of frame, and randomly occluding them, to avoid overfitting particular sparse depth configurations. We further push their limits in Sec. H of the Supp. Mat., where we test them on VOID150 with a 10× reduction in sparse points from the training set (VOID1500) and observe the same trend of improvements. This shows that AugUndo significantly improves robustness of models to changes in sparse depth.

Zero-shot generalization. We test depth completion models trained on VOID on NYUv2 and ScanNet. Tab. 3 shows an average of 23.2% improvement on NYUv2 and 27.6% on ScanNet. Applying AugUndo to VOICED greatly improve its generalizability to both NYUv2 and ScanNet. This is likely due to the scaffolding densification employed by VOICED, where the network can overfit to scaffoldings of particular sparse depth configurations and therefore, does not generalize well when presented with sparse depth maps with different configurations. Like our sensitivity study (Fig. 2, VOID500), occlusion and geometric augmentations introduce variation into the sparse depth configurations and densities, which alleviates overfitting to specific point clouds or 3D scenes; hence,

Table 3: *Zero-shot transfer from VOID to NYUv2 and ScanNet for depth completion.* AugUndo improves generalization of models trained on VOID to novel datasets by an average of 25.4% for all evaluation metrics (in millimeters) across both datasets.

	NYUv2				ScanNet			
Method	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓
VOICED	2240±143.90	2427±143.49	211±9.60	238±10.89	1562±136.79	1764±146.39	270±17.25	311±17.36
+ AugUndo	**990**±82.48	**1181**±65.55	**110**±6.11	**132**±6.01	**638**±44.74	**791**±79.48	**131**±7.68	**170**±6.32
FusionNet	132.24±2.12	236.16±4.59	28.68±0.42	61.87±1.20	109.47±3.01	206.33±6.11	55.45±1.56	122.52±2.04
+ AugUndo	**124.93**±3.05	**227.23**±4.96	**25.70**±0.41	**54.09**±0.87	**100.64**±2.31	**195.85**±5.85	**45.98**±0.78	**99.90**±5.00
KBNet	138.31±5.74	257.99±10.36	25.48±0.63	51.77±0.99	103.05±4.99	217.12±13.35	36.23±1.12	76.55±2.90
+ AugUndo	**118.60**±3.44	**231.13**±8.85	**22.06**±0.31	**47.07**±0.70	**82.53**±4.33	**175.30**±11.13	**29.87**±1.06	**63.78**±1.30

Table 4: *Zero-shot transfer from VOID to NYUv2 and ScanNet for depth estimation.* AugUndo consistently improves generalization (in meters) for monocular depth models.

Dataset	Method	MAE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
NYUv2	Monodepth2	0.432±0.003	0.556±0.005	0.205±0.001	0.159±0.001	0.683±0.005	0.907±0.001	0.975±0.001
	+ AugUndo	**0.415**±0.005	**0.537**±0.006	**0.196**±0.002	**0.148**±0.003	**0.700**±0.008	**0.915**±0.002	**0.977**±0.001
	HR-Depth	0.424±0.003	0.549±0.003	0.201±0.002	0.154±0.002	0.692±0.003	0.910±0.002	**0.976**±0.001
	+ AugUndo	**0.421**±0.009	**0.542**±0.011	**0.199**±0.005	**0.152**±0.006	**0.696**±0.009	**0.913**±0.004	**0.976**±0.001
	Lite-Mono	0.480±0.012	0.616±0.018	0.231±0.008	0.199±0.015	0.637±0.009	0.879±0.009	0.964±0.004
	+ AugUndo	**0.468**±0.003	**0.595**±0.003	**0.225**±0.004	**0.187**±0.005	**0.646**±0.002	**0.889**±0.003	**0.968**±0.001
ScanNet	Monodepth2	0.284±0.004	0.368±0.005	0.177±0.002	0.097±0.002	0.741±0.006	0.931±0.003	0.980±0.001
	+ AugUndo	**0.270**±0.003	**0.351**±0.004	**0.169**±0.001	**0.088**±0.002	**0.759**±0.004	**0.937**±0.001	**0.982**±0.001
	HR-Depth	0.282±0.004	0.366±0.005	0.175±0.002	0.095±0.002	0.743±0.005	0.929±0.003	0.979±0.002
	+ AugUndo	**0.274**±0.007	**0.357**±0.009	**0.172**±0.005	**0.092**±0.005	**0.754**±0.009	**0.935**±0.003	**0.981**±0.001
	Lite-Mono	**0.296**±0.002	0.388±0.007	**0.185**±0.003	0.109±0.006	**0.731**±0.002	0.921±0.002	0.976±0.001
	+ AugUndo	**0.296**±0.001	**0.382**±0.001	**0.185**±0.001	**0.105**±0.002	0.728±0.002	**0.924**±0.001	**0.977**±0.001

improving VOICED from 49.9% and 52.6% on NYUv2 and ScanNet, respectively. For FusionNet and KBNet, we still see nontrivial improvement: FusionNet improves by 8% and 12.1% on NYUv2 and ScanNet, and KBNet by 11.7% and 18.3%. This further validates our hypothesis that by applying a more diverse set of transformations, we are able to improve generalization to new datasets.

For depth estimation, we tested Monodepth2, HR-Depth, Lite-Mono (trained on VOID) for zero-shot generalizability to NYUv2 and ScanNet. The results are shown in Tab. 4, where training with AugUndo consistently improves generalization errors across all evaluation metrics for both NYUv2 and ScanNet. Notably, for RMSE metric, Monodepth2 improves from 0.556 to 0.537, HR-Depth from 0.549 to 0.542, Lite-Mono from 0.616 to 0.595. For ScanNet, similar improvement on RMSE can also be observed, where Monodepth2 improves from 0.368 to 0.351, HR-Depth from 0.366 to 0.357, and Lite-Mono from 0.388 to 0.382. The improvement in RMSE metric, which is sensitive to outliers, highlights AugUndo's ability to model different input data distributions, collected by a different camera of different orientation with different object colors and layouts.

Results on KITTI. We begin with quantitative results for depth completion. While AugUndo consistently improves all methods (Tab. 5), we note that the improvement is less, but respectable, in this case: ≈5.2% overall with the largest gain in FusionNet of 6.41%. This is largely due to the small scene varia-

Table 5: *Depth completion on KITTI.* AugUndo consistently improves all methods for all metrics (in millimeters) by approximately 5.2% on average.

Method	MAE ↓	RMSE ↓	iMAE ↓	iRMSE ↓
VOICED	318.59±7.74	1,213.60±17.49	1.30±0.05	3.72±0.04
+ AugUndo	**295.41**±0.30	**1,159.27**±5.44	**1.20**±0.02	**3.49**±0.03
FusionNet	285.55±2.16	1,174.47±10.67	1.20±0.03	3.45±0.08
+ AugUndo	**267.69**±1.85	**1,157.07**±4.61	**1.08**±0.02	**3.19**±0.03
KBNet	263.90±3.63	1,130.66±6.22	1.05±0.01	3.24±0.04
+ AugUndo	**256.37**±1.00	**1,114.53**±3.79	**1.01**±0.01	**3.13**±0.03

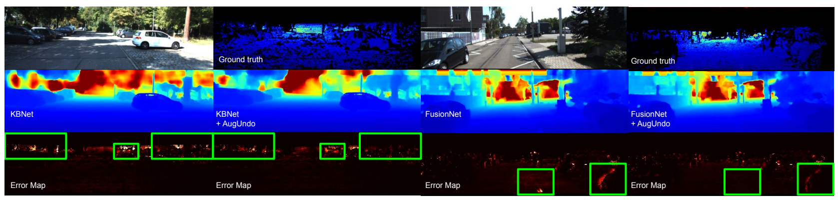

Fig. 3: *Depth completion on KITTI.* We compare KBNet and FusionNet trained with standard augmentations and with AugUndo. AugUndo consistently produces lower errors in highlighted regions where structures may have arbitrary orientation (vegetation) and regions near image border that typically lacks correspondence during training.

tions in the outdoor driving scenarios, i.e., ground plane with vehicles, buildings on the sides, horizontal lidar scans, and largely planar motion. The dataset bias is strong enough to render vertical flip detrimental to performance. This is also evident in existing works as none utilizes vertical flip as augmentation. Nonetheless, AugUndo still improves performance, where point removal models different lidar returns patterns, and resizing simulates large variations in scales of objects observed in road scenes. Similar to indoors, translations help model occlusions by shifting the projection with different principal points. Fig. 3 compares FusionNet and KBNet trained using conventional augmentations and AugUndo. AugUndo improves over objects that may have diverse orientations (i.e., tree branches and vegetation), thanks to rotation augmentations. Improvements are also observed in "small" objects at far regions where resizing can simulate zooming in/out to emulate objects of different sizes. Through translation and occlusion, AugUndo also improves on occlusion boundaries during training (highlighted), making estimates near image borders and object boundaries more robust.

We additionally show results for monocular depth estimation on KITTI in Tab. 6, where we compare the standard augmentation pipelines used by [15, 35, 81] and AugUndo. We observe similar trends in performance gain as we did in depth completion trained on KITTI: Applying our set of augmentations improves most metrics for all methods. We observe notable improvements in $\delta < 1.25$, the most difficult accuracy metric, from 0.869 to 0.879 in Monodepth2, from 0.879 to 0.883 in HR-Depth, and 0.862 to 0.863 in Lite-Mono. For Monodepth2, we also observe a large reduction in AbsRel error, improving it from 0.117 to 0.113.

Table 6: *Monocular depth estimation on KITTI.* AugUndo consistently improve on evaluation metrics (in meters) across different models.

Method	MAE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2	2.315±0.005	4.794±0.035	0.117±0.001	**0.845**±0.030	0.869±0.004	0.959±0.001	**0.982**±0.001
+ AugUndo	**2.237**±0.014	**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	2.226±0.004	4.626±0.032	0.113±0.001	0.797±0.022	0.879±0.002	0.961±0.001	0.982±0.000
+AugUndo	**2.185**±0.013	**4.610**±0.029	**0.111**±0.001	**0.794**±0.021	**0.883**±0.001	**0.962**±0.001	**0.983**±0.001
Lite-Mono	2.338±0.005	4.821±0.027	0.121±0.001	0.875±0.007	0.862±0.002	**0.955**±0.001	0.980±0.000
+ AugUndo	**2.314**±0.030	**4.780**±0.055	**0.120**±0.001	**0.849**±0.019	**0.863**±0.004	**0.955**±0.001	**0.981**±0.001

Similarly, for Lite-Mono, we also boosted performance across all metrics, with a particularly high reduction in SqRel from 0.875 to 0.849.

Remarks. As AugUndo only augments the inputs and modifies the loss computations, which add negligible time to training, we note that the performance improvements from it are obtained nearly for free. Yet, the percentage gain, however, is similar to improvements obtained by each successive state-of-the-art: For example, the improvement of Lite-Mono over HR-Depth , and that of HR-Depth over PackNet for monocular depth estimation; similarly, the improvement of FusionNet over VOICED and KBNet over FusionNet for depth completion.

5 Discussion

Conventionally, data augmentation aims to seek visual invariance and create a collection of equivalent classes, i.e., identifying an image and its augmented variant as the same. For geometric tasks, the underlying equivalence is in the 3D scene structures under various illumination conditions, camera, viewpoints, occlusion, etc. Assuming a rigid scene, the shapes populating it should persist regardless of the nuisance variables. This motivates the use of geometric augmentations, as it simulates the nuisances within the image. However, adoption of geometric augmentations in unsupervised geometric tasks are obstructed by artifacts introduced during transformations (see Sec. D of Supp. Mat.). AugUndo lifts this obstacle by "undo-ing" the augmentations before computing the loss.

While AugUndo enables scaling up augmentations for unsupervised training, i.e., depth completion, it may also be applicable for supervised methods; though, we posit that the gain to be less as the artifacts induced from photometric and geometric augmentations of an image tend to be larger than those of a piecewise smooth depth map. AugUndo is also limited to 2D augmentations; whereas, nuisances modeled by it are projections of the 3D scene (see limitations in Sec. M in Supp. Mat.). We leave extensions to 3D for future work. We also only consider a single image and sparse depth map as input. Likewise, extensions can be made towards multi-frame tasks such as stereo, optical flow, pose estimation, etc., but one must account for their specifics and problem setups, i.e., stereo assumes frontoparallel views. This is outside of our scope, so we leave them as future directions. This work paves way for the empirical success in unsupervised geometric tasks that we have observed in other visual recognition tasks.

Acknowledgements. This work was supported by was supported by NSF 2112562 Athena AI Institute.

References

1. Chen, Y., Schmid, C., Sminchisescu, C.: Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7063–7072 (2019)
2. 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)
3. 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)
4. Choi, H., Lee, H., Kim, S., Kim, S., Kim, S., Sohn, K., Min, D.: Adaptive confidence thresholding for monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12808–12818 (2021)
5. 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)
6. 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)
7. 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)
8. 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)
9. 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)
10. Fei, X., Wong, A., Soatto, S.: Geo-supervised visual depth prediction. IEEE Robotics and Automation Letters **4**(2), 1661–1668 (2019)
11. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
12. Garg, R., Bg, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 740–756. Springer (2016)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)

14. 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)
15. 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)
16. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2485–2494 (2020)
17. Guizilini, V., Hou, R., Li, J., Ambrus, R., Gaidon, A.: Semantically-guided representation learning for self-supervised monocular depth. In: International Conference on Learning Representations (2019)
18. Hu, M., Wang, S., Li, B., Ning, S., Fan, L., Gong, X.: Penet: Towards precise and efficient image guided depth completion. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13656–13662. IEEE (2021)
19. 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)
20. Jeon, J., Lim, H., Seo, D.U., Myung, H.: Struct-mdc: Mesh-refined unsupervised depth completion leveraging structural regularities from visual slam. IEEE Robotics and Automation Letters **7**(3), 6391–6398 (2022)
21. Kam, J., Kim, J., Kim, S., Park, J., Lee, S.: Costdcnet: Cost volume based depth completion for a single rgb-d image. In: European Conference on Computer Vision. pp. 257–274. Springer (2022)
22. Klingner, M., Termöhlen, J.A., Mikolajczyk, J., Fingscheidt, T.: Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 582–600. Springer (2020)
23. Krishna, S., Vandrotti, B.S.: Deepsmooth: Efficient and smooth depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3357–3366 (2023)
24. Kumar, V.R., Klingner, M., Yogamani, S., Milz, S., Fingscheidt, T., Mader, P.: Syndistnet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 61–71 (2021)
25. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016)
26. 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)
27. 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)
28. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurate o (n) solution to the pnp problem. International journal of computer vision **81**(2), 155 (2009)
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. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1119–1127 (2015)
31. Lin, Y., Cheng, T., Zhong, Q., Zhou, W., Yang, H.: Dynamic spatial propagation network for depth completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1638–1646 (2022)
32. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence **38**(10), 2024–2039 (2015)
33. 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)
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. 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)
36. 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)
37. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5667–5675 (2018)
38. 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)
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. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621 (2017)
41. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3227–3237 (2020)
42. Poggi, M., Tosi, F., Mattoccia, S.: Learning monocular depth estimation with unsupervised trinocular assumptions. In: 2018 International conference on 3d vision (3DV). pp. 324–333. IEEE (2018)
43. 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/CVF Conference on Computer Vision and Pattern Recognition. pp. 3313–3322 (2019)
44. Qu, C., Liu, W., Taylor, C.J.: Bayesian deep basis fitting for depth completion with uncertainty. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16147–16157 (2021)
45. 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)

46. 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)
47. 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)
48. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12240–12249 (2019)
49. 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)
50. 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)
51. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)
52. 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)
53. 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)
54. 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)
55. Taylor, L., Nitschke, G.: Improving deep learning with generic data augmentation. In: 2018 IEEE symposium series on computational intelligence (SSCI). pp. 1542–1547. IEEE (2018)
56. Tosi, F., Aleotti, F., Poggi, M., Mattoccia, S.: Learning monocular depth estimation infusing traditional stereo knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9799–9809 (2019)
57. 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)
58. 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)
59. 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)
60. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2022–2030 (2018)

61. Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: Self-supervised monocular depth hints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2162–2171 (2019)
62. Wong, A., Cicek, S., Soatto, S.: Targeted adversarial perturbations for monocular depth prediction. Advances in neural information processing systems **33**, 8486–8497 (2020)
63. 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)
64. 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)
65. 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)
66. 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)
67. 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)
68. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5354–5362 (2017)
69. 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/CVF International Conference on Computer Vision. pp. 2811–2820 (2019)
70. 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)
71. 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)
72. Yang, N., Stumberg, L.v., Wang, R., Cremers, D.: D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1281–1292 (2020)
73. 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)
74. Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R.: Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
75. Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R.: Lego: Learning edge with geometry all at once by watching videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 225–234 (2018)
76. 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)

77. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1983–1992 (2018)
78. Yu, Z., Sheng, Z., Zhou, Z., Luo, L., Cao, S.Y., Gu, H., Zhang, H., Shen, H.L.: Aggregating feature point cloud for depth completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8732–8743 (2023)
79. 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)
80. Zhan, H., Garg, R., Weerasekera, C.S., Li, K., Agarwal, H., Reid, I.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 340–349 (2018)
81. 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)
82. 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)
83. Zhang, Y., Guo, X., Poggi, M., Zhu, Z., Huang, G., Mattoccia, S.: Completionformer: Depth completion with convolutions and vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18527–18536 (2023)
84. Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., Mattoccia, S.: Monovit: Self-supervised monocular depth estimation with a vision transformer. In: 2022 International Conference on 3D Vision (3DV). pp. 668–678. IEEE (2022)
85. 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)
86. Zou, Y., Luo, Z., Huang, J.B.: Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 36–53 (2018)