An Adaptive Framework For Learning Unsupervised Depth Completion

Alex Wong[1], Xiaohan Fei[2], Byung-Woo Hong[3], and Stefano Soatto[1]

***Abstract*—We present a method to infer a dense depth map from a color image and associated sparse depth measurements. Our main contribution lies in the design of an annealing process for determining co-visibility (occlusions, disocclusions) and the degree of regularization to impose on the model. We show that regularization and co-visibility are related via the fitness (residual) of model to data and both can be unified into a single framework to improve the learning process. Our method is an adaptive weighting scheme that guides optimization by measuring the residual at each pixel location over each training step for (i) estimating a soft visibility mask and (ii) determining the amount of regularization. We demonstrate the effectiveness our method by applying it to several recent unsupervised depth completion methods and improving their performance on public benchmark datasets, without incurring additional trainable parameters or increase in inference time.**

***Index Terms*—Visual Learning, Sensor Fusion**

I. Introduction

INFERRING scene geometry from images supports a variety of tasks, from robotic navigation to image-based rendering. We focus on depth completion, the process of inferring a dense depth map at each instant of time, given an image and sparse depth measurements, which may be obtained from the same image(s) over time via structure-from-motion (SFM), or from a secondary sensor such as a lidar. This is an ill-posed problem, so the solution hinges on the choice of regularization or prior assumptions on the scene. The data fidelity criterion is the usual reprojection error customary in stereo and SFM and subject to visibility phenomena, occlusion and disocclusion. The regularizer imposes generic properties of the scene, for instance piece-wise smoothness and local connectivity.

There are two distinct phenomena where the data fidelity term (or reprojection error) does not meaningfully constrain the depth map to be inferred: Occlusions, and homogeneous regions. In the latter, there exists a wide range of disparities, one typically chooses the “simplest” as defined by the regularizer, for instance the smoothest depth map. In the former, no correct disparity map can fit the data term, since there is no displacement of one image that can match the other. Since no correct disparity exists, one should not penalize the reprojection error in the occluded regions, leaving the depth undefined. Both of these phenomena should be captured, ideally in a *unified fashion*. The main difference is that, whereas in homogeneous regions the data fidelity is already minimized, so the influence of the regularizer is increased automatically, in occluded regions the data fidelity term is uninformative and should be actively ignored so depth information should come from adjacent areas. Our goal is to devise an adaptive unsupervised learning framework that addresses both and fosters this process automatically.

The core of our approach is an adaptive weighting scheme that varies over space (image domain) and time (training steps) and informs (i) the probability of a given pixel being co-visible in two views (for weighting data fidelity) and (ii) the extent in which the prior assumptions (regularization) should be imposed – driven by the evidence in the data. To account for occlusions and disocclusions, we measure the fitness (residual) of the model to the data at each spatial position over each training time step. The result is a spatially varying soft visibility mask, relevant for spatial tasks such as navigation and manipulation, that *adapts* to the model over training time. The same residual can be used to determine the degree of regularization to impose on each spatial prediction, enabling a second set of adaptive weights. What makes this effective is the fact that, while the regularizers are generic (not informed by large image datasets), the way they are applied is driven by the evidence in the images, which leverages their strength (mostly simplicity) where appropriate, and limits the damage from their simplistic nature where necessary (e.g. across occluding boundaries). Together, the two sets of weights complement each other (i.e. occluded region requires regularization) and are combined into a single framework that can be generically applied to improve both *existing* and *yet-to-be-developed* unsupervised depth completion methods to guide their learning (optimization) to local minima that are more compatible with the data.

Counter to current trends, our framework requires *no extra trainable parameters*. It is entirely data-driven, leveraging information from the intermediate fitness between model and data as an adaptation signal for both sets of weights. It adaptively weights the data fidelity and regularization terms in the objective function during training and hence incurs *no additional run-time* during inference. Yet, our framework is able to consistently improve the performance of several

Manuscript received: October 15, 2020; Revised January 11, 2021; Accepted February 8, 2021 .

This paper was recommended for publication by Editor Cesar Cadena upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by ONR N00014-19-1-2229, ARO W911NF-17-1-0304, NRF-2017R1A2B4006023, and NRF-2018R1A4A1059731.

[1] Alex Wong, and Stefano Soatto are with Department of Computer Science, University of California, Los Angeles. Email: alexw@cs.ucla.edu, soatto@cs.ucla.edu

[2] Xiaohan Fei was with Department of Computer Science, University of California, Los Angeles when the work was conducted and is now with Amazon Web Services. Email: feixh@cs.ucla.edu

[3] Byung-Woo Hong is with the Department of Computer Science, Chung-Ang University, Korea. Email: hong@cau.ac.kr

Digital Object Identifier (DOI): see top of this page.

recent unsupervised depth completion algorithms across public benchmarks, such as KITTI [1] and VOID [2], and achieving new state-of-the-art – thus, demonstrating its effectiveness. To test the limits of our approach, we also provide a study on the model performance with lower density of the sparse points. Even with very few (0.05% density) points, our approach can *still* improve exisiting depth completion methods.

Our **contributions** are: (i) an annealed visibility mask that considers the fitness of model to data for determining and discounting occlusions during training, (ii) the use of residuals from multiple sensor modalities (image and depth) to determine the degree of regularization, and (iii) a unified framework that combines the adaptive weights for discounting occlusions and determining regularization where each set of weights plays a complementary role to the other; (iv) we show that our framework can be generically applied to unsupervised depth completion methods to achieve better performance without incurring additional trainable parameter or run-time complexity during inference.

II. Related Work

Supervised Depth Completion. Existing methods regress dense depth from an image and a sparse depth map by minimizing the difference between predictions and ground truth. [3] computed confidence from convolutions and propagates it through the layers. [4] performed upsampling followed by convolution to fill the missing values. [5], [6], [7] used two branches to process image and sparse depth separately. [5] used early fusion with a ResNet encoder, while [6] used late fusion. [7] also used late fusion, but with NASNet encoders and jointly learned depth and semantic segmentation. [8] proposed a 2D-3D fusion network. [9] learned confidence maps for guidance and [10], [11] also used surface normals. [12] formulated the problem as compressive sensing and [13] as morphological operators.

All of these methods are supervised. They require ground-truth, often unavailable, or the product of post-processing and aggregation over a number of consecutive frames [1]. Such supervision is not scalable; instead, we learn to predict dense depth by fusing information from the abundant un-annotated images and sparse depth data.

Unsupervised (Self-supervised) Depth Completion. Unsupervised methods learn depth by minimizing the discrepancy between prediction and sparse depth input, and between the given image and its reconstructions from additional (stereo or temporally adjacent) frames that are available only during training. Stereo methods [6], [14] predict disparity to reconstruct the given image from its stereo-counterpart and synthesize depth from focal length and baseline. These methods are generally limited to outdoor scenarios. Monocular methods [2], [5], [15] jointly learn depth and pose by projecting from temporally adjacent frames to a given image.

As depth completion is an ill-posed problem, regularization is needed. [2], [5], [14] used a generic local smoothness prior that is static with respect to the spatial domain of image and the temporal domain of optimization. Whereas, [6] utilized a *learned prior* (a separate network trained on ground-truth depth) to regularize predictions. [15] learned a topology prior on the sparse points from synthetic data and used it as regularization. We note that supervision from a network trained on a specific domain (e.g. outdoors) will not generalize (e.g. indoors) – defeating the purpose of unsupervised methods. Hence, we forgo the use of a learned prior, but instead propose a generic form of regularization that incorporates the local fitness of the current model estimate to data. Unlike conventional regularization, our approach is a locally adaptive, data-driven weighting scheme that varies in space and time and optimization to more desirable local minima.

Adaptive Weighting Schemes. Many imaging problems are cast into the optimization of an energy function that consists of data fidelity and regularization, where their relative significance is typically determined by a *static* scalar, which often leads to undesirable local minima due to *heteroscedasticity* of residual measuring a discrepancy between model and data. [16] determined the regularization parameter based on noise variance, and [17] on the cross-validation criterion. For depth completion, [2], [5], [14] determines the degree of regularization based on the the image gradient. However, this weighting scheme is still *static* with respect to a given image. [18], [19], [20] proposed adaptive regularization in the spatial domain and over the course of optimization based on the local residual. However, their method considers only a single frame. In contrast, we propose an adaptive data-driven algorithm that deals with multiple frames obtained from multiple sensor modalities (image and depth).

Unlike previous works, our method also considers occlusions and disocclusions in the data term. [2], [5], [6], [14] uniformly penalized all predictions without accounting for them. Unsupervised monocular depth prediction methods [21], [22], [23], [24] used an extra network to explicitly learn visibility masks by jointly minimizing an unsupervised photometric loss and a penalty for the cardinality of the mask (to avoid degenerate solution of all zeros). We discount unresolved residuals (due to visibility) over the course of optimization *without* incurring an extra network nor training time.

Uncertainty in Estimation. Our work is related to measuring uncertainty for 3D reconstruction. [25], [26] proposed to learn uncertainty from groundtruth for stereo. [27], [28] learned confidences based on deviation from median disparity. [29] did so in structure-from-motion (SfM) by leveraging existing SfM systems and [21], [22], [23] in monocular depth prediction. Unlike them, we showed that uncertainty or confidence does not need to be learned, but can be observed given the data. Our work is more in line with classic stereo works [30] in using the matching cost as a confidence measure, but unlike them, we used it to guide learning.

III. Motivation

Our goal is to recover a 3D scene from an RGB image $I : \Omega \subset \mathbb{R}^2 \mapsto \mathbb{R}^3_+$ and its associated sparse depth measures $z : \Omega_z \subset \Omega \mapsto \mathbb{R}_+$ in an unsupervised learning framework, where depth information is inferred by exploiting additional stereo imagery [6], [14] or temporally adjacent frames [2], [5] during the training phase. In this work, we assume that temporally adjacent frames, I_τ for $\tau \in \{-1, +1\}$ where I_{-1} denotes the previous frame and I_{+1} the next one with respect to I, are available. Thus, a training example comprises (I, I_τ, z). Note: our method can easily be extended to stereo training. To learn depth, unsupervised depth completion methods minimize a loss function $\mathcal{L}$ that mainly consists of data fidelity $\mathcal{D}$ and regularization $\mathcal{R}$ terms:

$$\mathcal{L}(\hat{z}) = \alpha\mathcal{D}(\hat{z}) + \gamma\mathcal{R}(\hat{z}), \quad \hat{z} = f(\theta; I, z), \tag{1}$$

where α and γ are pre-defined positive scalars that are applied *uniformly* to data fidelity and regularization terms to modulate their trade-off.

The model f, parameterized by θ, takes an image I and sparse depth z, which resides in $\Omega_z \subset \Omega$, as input and produces dense depth $\hat{z} : \Omega \mapsto \mathbb{R}_+$. To learn depth, we minimize Eqn. 1 over the entire training dataset. The data fidelity term $\mathcal{D}$ is designed to penalize the combination of discrepancies (i) between z and its prediction $\hat{z}$ and (ii) between I and its reconstruction $\hat{I}_\tau$. The reconstruction $\hat{I}_\tau$ from I is obtained by the following projection equation:

$$\hat{I}_\tau(x) = I_\tau\big(\mathrm{p}(\, g_\tau \, K^{-1} \begin{bmatrix} x \\ 1 \end{bmatrix} \hat{z}(x))\big), \tag{2}$$

where $x \in \Omega$, $\tau \in \{-1, +1\}$, g_τ is the relative pose between I and I_τ, K the camera intrinsics, and p the projection operation. There are two main problems in Eqn. 1: (i) Because $\mathcal{D}$ is subject to occlusions and disocclusions when registering I_τ to I and vice versa, occluded and disoccluded regions will yield high reconstruction errors (residuals) and a uniform weighting scheme α will penalize these regions despite the lack of co-visibility. (ii) Because $\mathcal{R}$ is commonly a local smoothness (e.g. total variation of $\hat{z}$) or a forward-backward consistency term, a uniform weighting scheme γ will bias $\hat{z}(x)$ for $x \in \Omega$ to be smooth or consistent with another prediction *without* considering the residuals or correctness of $\hat{z}(x)$, which can cause performance to degrade.

Hence, neither α nor γ should be static, but instead *adapt* to the model and data for each prediction $\hat{z}(x)$. As α and γ are both related to data fidelity residuals (which evolves throughout training), one must consider the temporal interplay between data fidelity and regularization over the course of optimization. Thus, we propose residual-guided adaptive weighting functions $\alpha_\tau(x)$ and $\gamma(x)$, that vary in both space (image domain) and time (optimization step), to determine visibility and regularization. We combine them into a simple yet effective framework (see Fig. 1), where their complementary effects (i.e. occluded regions require regularization) can improve baseline unsupervised depth completion algorithms without any additional trainable parameters.

IV. Determining Visibility Over Time

Given an image pair (I, I_τ) and the depth predictions $\hat{z}(x)$, the reconstruction $\hat{I}_\tau$ suffers from occlusions and disocclusions because I_τ is captured from a different viewpoint. A static (uniform weighting) α penalizes all discrepancies between $\hat{I}_\tau$ and I_τ equally regardless of visibility constraints, i.e. co-visibility, occlusion or disocclusion, and thus *requires* the model to resolve regions that are not co-visible.

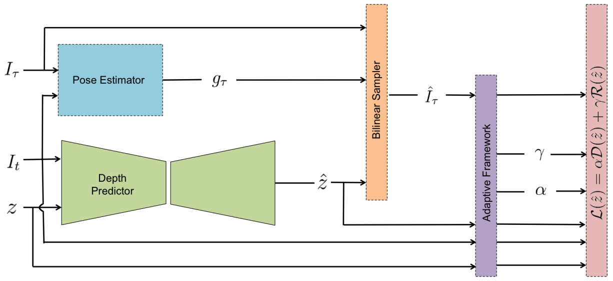

Fig. 1. *Diagram of the training pipeline using our framework.* Given the predicted depth $\hat{z}$, sparse depth z, image I_t and its reconstructions $\hat{I}_\tau$, our framework (purple) is comprised of α (consists of α_τ for $\tau \in \{-1, +1\}$) and γ for adaptively weighting the data fidelity $\mathcal{D}$ and regularization $\mathcal{R}$ in the loss function (red). Our framework does not require any additional trainable parameters nor additional run-time complexity during inference; the only component required during inference is the depth predictor (green).

Let us consider a scenario where all co-visible correspondences are found, the reconstruction residual will still be non-zero and hence the gradients will continue to update the model parameters θ to find unresolvable correspondences up to the allowed regularization, causing the model to *move away* from the desired solution. One may discount occlusions and disocclusions with a binary mask based on a fixed threshold (i.e. in traditional SFM, stereo). This is applicable at convergence, when all correspondences have been found. However, at early time steps, predictions are largely random and hence will yield high residuals. Thresholding would discount the training signal and in turn *impede learning*. Hence, an adaptive weighting scheme $\alpha_\tau \in [0, 1]$ for (I, I_τ) should weight all pixels equally at the early stages of training. As the model becomes more confident in the correspondences found over the course of training, $\alpha_\tau \rightarrow 0$ for regions with high residuals, gradually discounting the errors.

A. *Residual Function*

We begin with a simple residual function as a measure for determining whether a pixel is co-visible, or occluded or disoccluded. Assuming images with intensity range of $[0, 1]$:

$$\delta_\tau(x) = |I(x) - \hat{I}_\tau(x)| \text{ for } x \in \Omega \tag{3}$$

measures the discrepancy between I, and its reconstruction $\hat{I}_\tau$ (photometric error). Note: δ_τ can be replaced by a more sophisticated measure such as SSIM [31], but we aim to demonstrate the effectiveness of our proposed scheme with a simple one. We then normalize the residual δ_τ to have a zero-mean distribution with unit variance.

$$\mu_\tau = \frac{1}{|\Omega|}\sum_{x\in\Omega}\delta_\tau(x), \quad \sigma_\tau^2 = \frac{1}{|\Omega|}\sum_{x\in\Omega}(\delta_\tau(x) - \mu_\tau)^2, \tag{4}$$

$$\rho_\tau(x) = \frac{\delta_\tau(x) - \mu_\tau}{\sqrt{\sigma_\tau^2 + \epsilon}}, \tag{5}$$

where $x \in \Omega$ and ϵ is a small positive scalar used for numerical stability. In the next section, we will use ρ_τ as a cue to determine if a pixel is co-visible, or occluded or disoccluded by constructing a soft visibility mask α_τ that evolves over training time.

B. *Discounting Occlusions and Disocclusions*

The weighting function α_τ assigns the probability of co-visibility between I and $\hat{I}_\tau$ for each pixel by adaptively

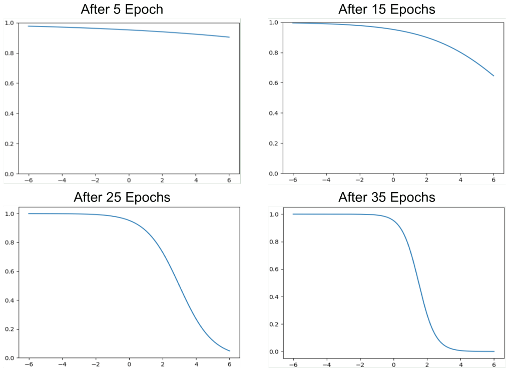

Fig. 2. *The shape of $\alpha_\mathcal{T}$ from mean residual $\mu_\mathcal{T}$ sampled after the 5^{th}, 15^{th}, 25^{nd} and 35^{th} epoch.* x-axis denotes $\rho_\mathcal{T}(x)$ and y-axis denotes the value of $\alpha_\mathcal{T}$. $\alpha_\mathcal{T}$ begins at a flat curve close to 1 and over time sharpens into a flipped sigmoid. Binary thresholding is a special case of our approach.

adjusting a flipped sigmoid based on $\rho_\mathcal{T}$ for every time step (Fig. 2). Co-visible pixels will have a higher weight, while occluded or dis-occluded pixels will have a lower weight:

$$\alpha_\mathcal{T}(x) = 1 - \frac{1}{1 + \exp(-(a\rho_\mathcal{T}(x) - b))}, \tag{6}$$

where $a > 0$ controls the curvature (steepness) of the sigmoid and $b \geq 0$ the shift. To enable adaptation over training time, we vary a and b based on the mean residual $\mu_\mathcal{T} \in [0, 1]$. The steepness parameter a is designed to gradually increase over training as the overall residual decreases:

$$a = \frac{a_0}{\mu_\mathcal{T} + \epsilon} \tag{7}$$

where a_0 is a positive scalar based on the range of image intensity. As we are unsure of the correspondences during the early stages of training, $\alpha_\mathcal{T}$ should be uniform over the spatial domain Ω, which occurs as $a \to 0$. Towards convergence, $\alpha_\mathcal{T}$ takes on the shape of a flipped sigmoid to discount occlusions and disocclusions. Hence, we let a be inversely proportional to the mean residual $\mu_\mathcal{T}$ and we choose a_0 to be close to 0. At the start of training, $\mu_\mathcal{T}$ is large (making a small) and $\alpha_\mathcal{T}$ tends to a flat curve. As we converge, $\mu_\mathcal{T} \to 0$, making a large and giving $\alpha_\mathcal{T}$ sharper curvature.

Similarly, we also allow the shift parameter b of $\alpha_\mathcal{T}$ to vary over training time by making it a function of $\mu_\mathcal{T}$:

$$b = b_0(1 - \cos(\pi\, \mu_\mathcal{T})), \tag{8}$$

where b_0 is a positive constant used as the upper bound of the shift and $\mu_\mathcal{T} \in [0, 1]$ leading to $b \in [0, 2b_0]$, following a cosine decay rate. At the early time steps, $\mu_\mathcal{T}$ is large, and thus $b \to 2b_0$ causing $\alpha_\mathcal{T}$ to tend to 1. As residuals decrease over training time, $b \to 0$, resulting in $\alpha_\mathcal{T}$ being a centered flipped sigmoid function.

By making a and b a function of the mean residual $\mu_\mathcal{T}$, the weighting function $\alpha_\mathcal{T}$ becomes an annealing process to detect occlusions or disocclusions. Because $\alpha_\mathcal{T}$ is modulated by both the local (per-pixel) residual as well as the mean residual (generally decreases throughout training), $\alpha_\mathcal{T}$ will vary over both the image spatial domain and training time. For every $x \in \Omega$, $\alpha_\mathcal{T}(x) \approx 1$ at the early stage of the training, whereas $\alpha_\mathcal{T}(x)$ approaches either 0 or 1 towards the convergence of the training. We note that the binary mask produced by thresholding is a special case of our method with specific a and b. We construct Fig. 2 by sampling $\mu_\mathcal{T}$ over the course of training to illustrates how $\alpha_\mathcal{T}$ is guided by mean residual and varies over training time. Fig. 3 shows $\alpha_\mathcal{T}$ as an image. The co-visible pixels (yellow) are assigned higher weight; whereas, the occluded and disoccluded ones (blue) are assigned lower weight – as we train, the weight of those regions decreases as we are more confident in our predictions. In the data fidelity term $\mathcal{D}$, one can apply $\alpha_\mathcal{T}$ to the photometric error $\mathcal{D}_{ph}$ simply by:

$$\mathcal{D}_{ph}(\hat{z}) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \alpha_\mathcal{T}(x) |I(x) - \hat{I}_\mathcal{T}(x)|. \tag{9}$$

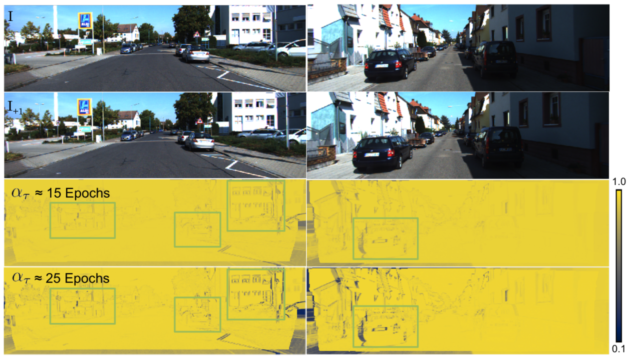
I
I+1
$\alpha_\mathcal{T}$ ≈ 15 Epochs
$\alpha_\mathcal{T}$ ≈ 25 Epochs
1.0
0.1

Fig. 3. *$\alpha_\mathcal{T}$ over training time.* $\alpha_\mathcal{T}$ varies spatially and over training time, and reduces weight of occlusions and disocclusions regions (e.g. the borders of the image and regions highlighted in green) as we become more confident in correspondences between I and I_{+1}.

V. Adaptive Regularization

Regularization is typically imposed uniformly over the prediction to make the depth completion problem well-posed. For instance, a local smoothness term assumes a smooth transition in $\hat{z}(x)$ and penalizes discontinuities, but does not account for object boundaries where depth discontinuities generally occur. Hence, uniformly imposing regularity may lead to an undesirably biased model (e.g. over-smoothing). To allow discontinuities along object boundaries, previous works, including, but not limited to [2], [5], [14], "adapt" to the data by weighting $\mathcal{R}(x)$ based on the image gradients $\nabla I(x)$ – reducing $\gamma(x)$, the regularization parameter, in textured regions. However, $\gamma(x)$ is *still static* with respect to the image (same weights for the same image). Also, this does not consider *residuals* where regularization not only propagates the incorrect solution, but also restricts the model from exploring the solution space (i.e. predicting large disparities). This also holds for other regularizers, such as temporal consistency; enforcing consistency with incorrect predictions only introduces more errors.

Hence, $\gamma \in [0, 1]$ should adaptively imposes regularization based on residuals from both image and sparse depth. $\gamma(x)$ follows two simple principles for a given $\hat{z}(x)$: (i) the higher the residual, the lower the regularity. This not only lowers the influence of incorrect predictions in a local neighborhood, but also gives a model the flexibility to maximize its fitness to data. (ii) the earlier the time step, the lower the regularity. The local

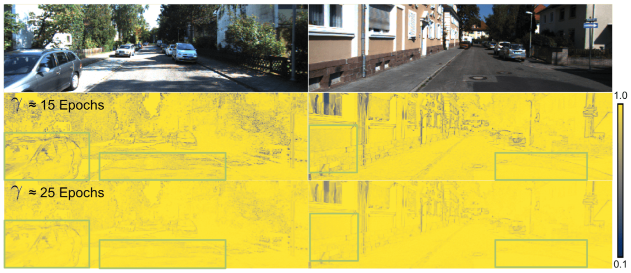

Fig. 4. *γ over training time.* γ is low during early stages of training and increases over time as mean residuals decrease. Regions of small residuals due to noise and slight illumination change (highlighted) are gradually regularized. For a local smoothness term, γ first allow the model to search for better correspondences and gradually impose smoothness.

residual at early steps can be low depending on initialization, applying regularization effectively limits the scope of the solution space. Hence, small amounts of regularity should be imposed to allow the model to explore. To illustrate these seemingly counter-intuitive principles, let's consider stereo matching. One can predicted disparity up to the amount allowable by regularization. If $\gamma(x)$ is large, then one cannot find long range correspondences; hence, we want to reduce $\gamma(x)$. Once the correspondence is found, we can leverage the correct prediction to inform its neighbors' predictions (e.g. local smoothness).

A. Residual Functions

We will reuse the image reconstruction residual (Eqn. 3) as our adaptation signal from an image. As we assume that there are two temporally adjacent frames, I_τ for $\tau \in \{-1, +1\}$, there exists two reconstructions of I_τ to guide γ. Following our first principle to apply regularization when residual is low, for each $x \in \Omega$, we choose the minimum residual of the two reconstructions:

$$\delta_i(x) = \min_{\tau}(\delta_\tau(x)) \text{ for } x \in \Omega. \tag{10}$$

To obtain an adaptation signal from depth input, we consider the sparse depth reconstruction residual:

$$\delta_z(x) = \begin{cases} |\hat{z}(x) - z(x)|, & \text{if } x \in \Omega_z, \\ 0, & \text{if } x \in \Omega \backslash \Omega_z. \end{cases} \tag{11}$$

Next, we will use δ_i and δ_z to construct γ_i (from image) and γ_z (from sparse depth), and combine them to form our adaptive regularization weighting scheme γ.

B. Image and Sparse Depth as Guidance

To realize our second principle of having a small $\gamma(x)$ at early time steps, we note a keen observation: while some local residuals may be small, the mean residual will be large and will gradually decrease over the course of optimization – making it a good proxy for training time. Hence, $\gamma(x)$ should be inversely proportional to the mean residual. First, we model γ_i, adaptive weights guided by image residuals, with a negative exponential function:

$$\gamma_i(x) = \exp(-c_i\, \mu_i\, \delta_i(x)) \text{ for } x \in \Omega \tag{12}$$

$$\mu_i = \frac{1}{|\Omega|} \sum_{x \in \Omega} \delta_i(x) \tag{13}$$

where c_i is a positive scalar based on the range of image intensities. Similarly, we also construct γ_z, the set of adaptive weights from sparse depth residuals:

$$\gamma_z(x) = \exp(-c_z\, \mu_z\, \delta_z(x)) \text{ for } x \in \Omega_z \tag{14}$$

$$\mu_z = \frac{1}{|\Omega_z|} \sum_{x \in \Omega_z} \delta_z(x) \tag{15}$$

where c_z is a positive scalar based on the range of depth measurements. Both γ_i and γ_z are modulated by their respective local and mean residuals. At early steps, both are low and increase over time, except where $\hat{z}(x)$ yield high residuals. We note that modulating γ_i and γ_z with their mean residuals as a proxy of training time is more stable than using discrete training steps, which have no upper bound. If γ_i and γ_z directly depend on training steps, then they may modify the model even after convergence, and introduce instability. In contrast, the mean residual stays approximately constant at convergence and γ_i and γ_z will like-wise be stable.

Lastly, as noted by previous works, sparse depth and image may conflict due to noise in depth sensor, and illumination changes in images. To combine γ_i and γ_z, we assume depth measurements (when available) are more reliable and choose γ_z over γ_i, yielding the final adaptive weights:

$$\gamma(x) = \begin{cases} \gamma_z(x), & \text{if } x \in \Omega_z, \\ \gamma_i(x), & \text{if } x \in \Omega \backslash \Omega_z. \end{cases} \tag{16}$$

The behavior of γ is similar to anisotrophic diffusion at convergence since the regions of high residuals will be occlusion or disocclusions, which generally occurs across object boundaries (see Fig. 4). However, unlike Sec. IV-B, we chose a negative exponential over a sigmoid function because the negative exponential is less aggressive at convergence. Recall that α_τ approaches a binary mask, but we still need some regularity since the problem is ill-posed.

Assuming local smoothness as $\mathcal{R}$, one can apply γ by:

$$\mathcal{R}(\hat{z}) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \gamma(x) ||\nabla \hat{z}(x)||^2 \tag{17}$$

Together, α_τ and γ *complement each other*. During early time steps, α_τ is high and γ is low, allowing the model to explore the solution space for better correspondences. As residuals decrease over time, α_τ discovers occlusions or disocclusions and discounts them. This is precisely when we need regularity and consequently γ increases (see Fig. 3, 4).

We note that α_τ and γ are general and can be constructed with a stereo pair as well. In this case, there is only one reconstruction from a stereo-counterpart, $\delta_i(x)$ is simply the reconstruction residual (Eqn. 3) instead of the minimum residual from multiple views (Eqn. 10). To show that our framework is applicable to both stereo and monocular training paradigms, we use [14] as a baseline and construct α_τ and γ_i using stereo pairs (see Table II).

VI. Implementation Details

All models using our framework are trained from scratch. Our framework consists of computationally cheap operations

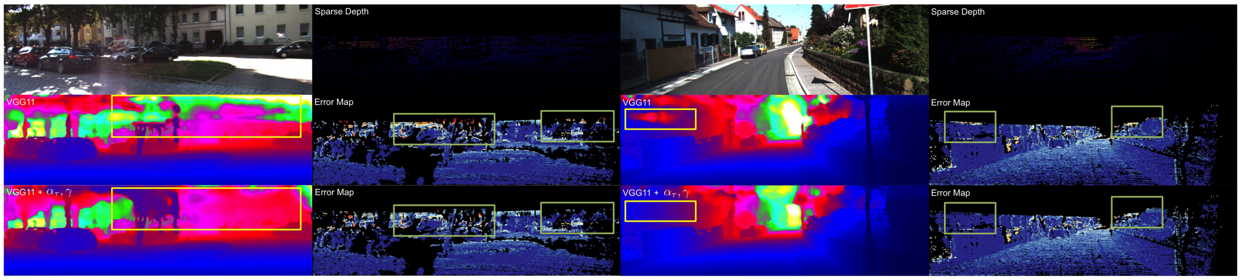

Fig. 5. *KITTI depth completion test set.* VGG11 completely missed the building, and tree on the left and the wall on the right (highlighted in yellow). By considering the fitness of the model to the data, our framework enables VGG11 to recover all of them. Green boxes highlight error regions for comparison.

TABLE I
ERROR METRICS

Metric	Definition
MAE	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{z}(x) - z_{gt}(x)\rvert$
RMSE	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{z}(x) - z_{gt}(x)\rvert^2\right)^{1/2}$
iMAE	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{z}(x) - 1/z_{gt}(x)\rvert$
iRMSE	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert 1/\hat{z}(x) - 1/z_{gt}(x)\rvert^2\right)^{1/2}$

Error metrics used in Table II, III. z_{gt} denotes the ground truth.

and only increases training time by $\approx 2.2\%$, and incurs no additional parameters or inference time.

Hyper-parameters: α_τ and γ are set based on the range of input. Image intensities are scaled between 0 and 1 for both KITTI and VOID. Depth ranges from 1m to 100m in KITTI and 0.1m to 10m in VOID. We choose $a_0 = 0.10$, $b_0 = 4.0$ and $\epsilon = 10^{-8}$ for α_τ. The same a_0 and b_0 are used for both α_τ. We set $c_i = 1.0$, $c_z = 0.01$ for γ to adjust for the difference in magnitude between image and depth values. The same hyper-parameters are used for all methods for both KITTI and VOID except for c_i, which we set to 0.70 (less aggressive weighting) for VOID since indoor scenes contains more textureless surfaces and requiring more regularity. For the same reason, to train [5] on VOID, we set the weight of their smoothness term to 1.0 (10× their proposed weight).

VII. Experiments and Results

We applied our adaptive framework (α_τ and γ) to recent unsupervised depth completion methods and evaluate the relative improvements on the KITTI [1] in Sec. VII-A (outdoors) and VOID [2] in Sec. VII-B (indoors).

A. KITTI Unsupervised Benchmark

KITTI provides $\approx 80,000$ synchronized stereo pairs and sparse depth maps of $\approx 1242 \times 375$ resolution for outdoor driving scenes. The sparse depth maps are captured by a Velodyne lidar sensor ($\approx$ 5% density) and projected onto the image frame. The ground-truth depth map is created by accumulating the neighbouring 11 raw lidar scans, with dense depth corresponding to the lower 30% of the images.

We apply our framework to [5], [14], and VGG8, and VGG11 of [2] and evaluate them on the KITTI validation set in Table II using error metrics in Table I. Due to the limit of one entry per method on the KITTI online test benchmark, we chose to show the relative improvements on the validation set – before and after applying α_τ and γ. The results listed are taken directly from their papers except for [14], which were not reported. We trained their model from scratch in Table II. While we primarily focus on the monocular training, we include [14] to show that our framework can also be applied to and improve methods using stereo training.

TABLE II
QUANTITATIVE RESULTS ON KITTI VALIDATION SET

Method	MAE	RMSE	iMAE	iRMSE
Ma	358.92	1384.85	1.60	4.32
Ma + our α_τ, γ	**332.54**	**1301.42**	**1.43**	**4.01**
Shivakumar	396.43	1285.79	1.37	4.05
Shivakumar + α_τ, γ	**346.18**	**1231.06**	**1.31**	**3.84**
VGG8 (Wong)	308.81	1230.85	1.29	3.84
VGG8 (Wong) + α_τ, γ	**298.89**	**1189.43**	**1.18**	**3.64**
VGG11 (Wong)	305.06	1239.06	1.21	3.71
VGG11 (Wong) + α_τ, γ	**291.57**	**1186.07**	**1.16**	**3.58**

Results of [5], [2] are taken from their papers. Results of [14] were not available; hence, we train [14] from scratch. Our approach (entries with + α_τ, γ) consistently improves all methods across all metrics.

Table II shows that our framework consistently improves all methods across all metrics. While our method can be used to improve both existing and yet-to-be-developed methods, the real test is whether it can boost an underperforming method over the state of the art. Hence, a **key comparison** is between VGG8, VGG8 + α_τ, γ and VGG11. Indeed, our framework improves an inferior method, VGG8, over the state-of-the-art VGG11 across all metrics on the KITTI depth completion validation set as well as the official online KITTI depth completion test set (Table IV, Supp. Mat.) to achieve the state-of-the-art on unsupervised depth completion. Fig. 5 shows that our framework can help VGG8 more correctly recover the scene. We note that the performance boost is *almost free* ($\approx +2.2\%$ in training time) – there is no additional parameters, pre- or post-processing, nor increase in inference time. The gain is solely from guiding the learning (optimization) process via adaptively weighting their objective function.

B. VOID Unsupervised Benchmark

VOID provides $\approx 47,000$ synchronized images and sparse depth maps of 640×480 resolution of indoor scenes. Sparse depth (≈ 1500 points, covering $\approx 0.5\%$ of the image) are the set of features tracked by XIVO [32]. The ground-truth depth maps are dense and are acquired by active stereo. The testing set contains 800 frames.

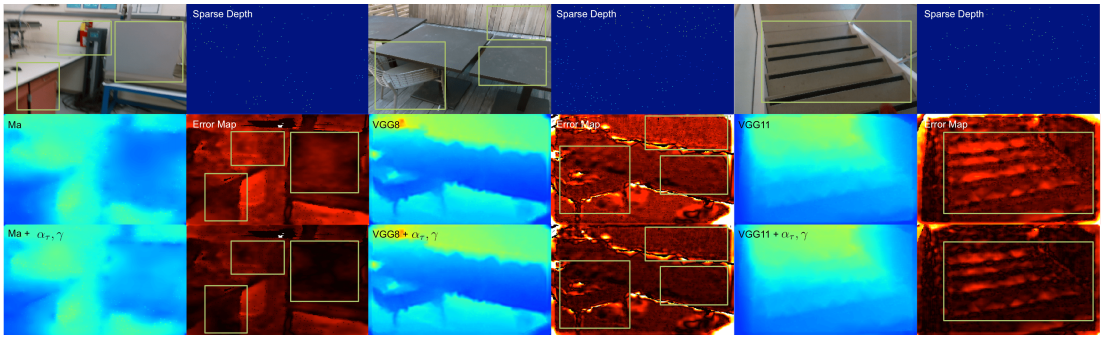
Sparse Depth
Sparse Depth
Sparse Depth
Ma
Error Map
VGG8
Error Map
VGG11
Error Map
Ma + α_τ, γ
VGG8 + α_τ, γ
VGG11 + α_τ, γ

Fig. 6. *VOID depth completion test set*. We apply our framework to [5], VGG8, and VGG11 of [2]. Our approach consistently improves the overall scene. While depth maps may look similar, the error map of each method using our framework (+ α_τ, γ) is a shade of red darker (lower error). For example, We observe large improvements in smooth surfaces (tables, walls, staircases) Green boxes highlight regions for comparison.

TABLE III
VOID TEST SET AND ABLATION STUDY ON α_τ AND γ

Method	MAE	RMSE	iMAE	iRMSE
Ma	178.85	243.84	80.12	107.69
Ma + α_τ, γ	**154.48**	**220.63**	**64.68**	**91.77**
VGG8 (Wong)	98.45	169.17	57.22	115.33
VGG8 (Wong) + α_τ, γ	**86.25**	**153.05**	**49.26**	**94.74**
VGG11 (Wong)	85.05	169.79	48.92	104.02
VGG11 (Wong) + α_τ	83.24	139.52	47.51	83.69
VGG11 (Wong) + γ	**78.20**	140.86	45.41	85.20
VGG11 (Wong) + α_τ, γ	78.79	**135.93**	**43.62**	**78.22**

Our framework (+ α_τ, γ) consistently improves all methods across all metrics. α_τ and γ provide complementary benefits. The ablation study (last 4 rows) on VGG11 shows that γ improves MAE and iMAE, and α_τ, RMSE and iRMSE. When used together, they achieve the best results..

TABLE IV
ABLATION STUDY OF VARIOUS DENSITY LEVELS ON VOID TEST SET

Method	MAE	RMSE	iMAE	iRMSE
		≈0.50% density		
VGG11 [2]	85.05	169.79	48.92	104.02
VGG11 + α_τ, γ	**78.79**	**135.93**	**43.62**	**78.22**
		≈0.15% density		
VGG11 [2]	124.11	217.43	66.95	121.23
VGG11 + α_τ, γ	**112.31**	**188.60**	**59.47**	**101.26**
		≈0.05% density		
VGG11 [2]	179.66	281.09	95.27	151.66
VGG11 + α_τ, γ	**155.01**	**262.54**	**83.55**	**140.98**

The percent density levels correspond to roughly 1500, 500, and 150 points, respectively. By applying our framework to VGG11 [2], we improve their model across all metrics and consistently across all density levels. These are the scenarios where unsupervised depth completion methods must rely on the image – due to the lack of sparse points. They are also the scenarios where our framework can provide improvements, especially for indoor scenarios. Thus, even at ≈0.05% density, we still boost performance.

We apply our framework on [5], VGG8 and VGG11. Since [5] and VGG8 did not report results on VOID, we trained their models from scratch. Note: [14] requires stereo pairs for training, so we cannot train their model on VOID.

VOID consists of indoor scenes with many textureless surfaces (e.g. walls, cabinets) and non-trivial 6 degrees of freedom motion. Hence, (i) regularization is even more important as the data fidelity term does not provide useful local information. This is where γ is helpful. By adjusting the regularity based on residuals, γ allows the model to find correspondences first, then impose regularization. Moreover, (ii) due to the large motion, occlusions and disocclusions can easily cause the model to leave a desirable local minimum. α_τ mitigates their impact by discounting them over time. The effectiveness of our framework can be seen in Table III and Fig. 6, where we improved all methods by large margins across all metrics. We hypothesize the large gain in iMAE and iRMSE metrics may be due to the low density of depth measurements. Hence, the model must rely heavily on the signal from the image, which is guided by α_τ and γ.

As an ablation study, we examine α_τ and γ (Table III) individually on VGG11 and find that both provide complementary benefits. γ provides more improvements to MAE and iMAE while α_τ improves RMSE and iRMSE. Note: the MAE improvement from γ (last 2 rows) is comparable to our full model. This is because α_τ reduces outliers (as measured by RMSE metrics) caused by occlusions and disocclusions while γ improves the overall accuracy of the scene (as measured by MAE metrics) through regularization.

To evaluate the effect of different density levels, we provide an ablation study on the VOID [2] dataset, which provides three levels: ≈0.50%, ≈0.15% and ≈0.05% of the image space – each of these densities corresponds to roughly 1500, 500, and 150 points. In Table IV, we show the results of VGG11 [2], directly taken from their paper, and the results of VGG11 trained with our framework. We observe consistent improve across all metrics and across all density levels. As the density of the input sparse depth decreases, one *must* rely on the image even more. For indoor, this becomes difficult as surfaces are commonly textureless and motion is more challenging (causing occlusions and dis-occlusions).

This is precisely where our method can provide improvements. Our framework produces a soft visibility mask α_τ to deal with occlusions and dis-occlusions and γ to determine the strength of regularization, which, in this setting, generally involves local smoothness and forward-backward consistencies. α_τ discounts occlusion and dis-occlusions over time so that the model does not get driven out of a desirable local minimum due to unresolvable residuals. In the case where correspondences are not found, γ allows the model enough flexibility to search for long-range matches (as opposed to uniform weight, which may restrict the model to shorter distances in the image space depending on selected scalar). Once a correspondence is found, γ increases regularization and propagates the solution to its neighbors, which directly impacts textureless regions, occluded and dis-occluded regions, and prevents over-smoothing. Together, α_τ and γ play complementary roles by discounting residuals at occluded and dis-occluded regions while propagating depth values from covisible regions to those locations.

VIII. Discussion

We have provided a general residual-driven framework for determining co-visibility and the degree of regularization over the optimization process. While our framework improves unsupervised depth completion methods without compromising run-time, it does require tuning several parameters depending on the range of sensors and environment. We use simple measure of residual and do not consider sparse depth in α_τ. We also assume depth measurements are reliable than images when constructing γ. In reality, both camera and depth sensors have failure modes. Perhaps considering hardware uncertainty can better combine the two. We leave this for future work. There is a long road ahead, but we hope that our simple framework can lay the foundation for better balancing of data fidelity and regularization through sensor fusion.

References

[1] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, “Sparsity invariant cnns,” in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.

[2] A. Wong, X. Fei, S. Tsuei, and S. Soatto, “Unsupervised depth completion from visual inertial odometry,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1899–1906, 2020.

[3] A. Eldesokey, M. Felsberg, and F. S. Khan, “Propagating confidences through cnns for sparse data regression,” in *Proceedings of British Machine Vision Conference (BMVC)*, 2018.

[4] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang, and H. Li, “Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion,” *IEEE Transactions on Image Processing*, vol. 29, pp. 3429–3441, 2019.

[5] F. Ma, G. V. Cavalheiro, and S. Karaman, “Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3288–3295.

[6] Y. Yang, A. Wong, and S. Soatto, “Dense depth posterior (ddp) from single image and sparse range,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3353–3362.

[7] M. Jaritz, R. De Charette, E. Wirbel, X. Perrotton, and F. Nashashibi, “Sparse and dense data with cnns: Depth completion and semantic segmentation,” in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 52–60.

[8] Y. Chen, B. Yang, M. Liang, and R. Urtasun, “Learning joint 2d-3d representations for depth completion,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 023–10 032.

[9] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, “Sparse and noisy lidar completion with rgb guidance and uncertainty,” in *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 2019, pp. 1–6.

[10] Y. Xu, X. Zhu, J. Shi, G. Zhang, H. Bao, and H. Li, “Depth completion from sparse lidar data with depth-normal constraints,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2811–2820.

[11] Y. Zhang and T. Funkhouser, “Deep depth completion of a single rgb-d image,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 175–185.

[12] N. Chodosh, C. Wang, and S. Lucey, “Deep Convolutional Compressed Sensing for LiDAR Depth Completion,” in *Asian Conference on Computer Vision (ACCV)*, 2018.

[13] M. Dimitrievski, P. Veelaert, and W. Philips, “Learning morphological operators for depth completion,” in *Advanced Concepts for Intelligent Vision Systems*, 2018.

[14] S. S. Shivakumar, T. Nguyen, I. D. Miller, S. W. Chen, V. Kumar, and C. J. Taylor, “Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion,” in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 13–20.

[15] A. Wong, S. Cicek, and S. Soatto, “Learning topology from synthetic data for unsupervised depth completion,” *IEEE Robotics and Automation Letters*, 2021.

[16] N. P. Galatsanos and A. K. Katsaggelos, “Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation,” *IEEE Transactions on image processing*, vol. 1, no. 3, pp. 322–336, 1992.

[17] N. Nguyen, P. Milanfar, and G. Golub, “Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement,” *IEEE Transactions on image processing*, vol. 10, no. 9, pp. 1299–1308, 2001.

[18] B.-W. Hong, J.-K. Koo, H. Dirks, and M. Burger, “Adaptive regularization in convex composite optimization for variational imaging problems,” in *German Conference on Pattern Recognition*. Springer, 2017, pp. 268–280.

[19] B.-W. Hong, J. Koo, M. Burger, and S. Soatto, “Adaptive regularization of some inverse problems in image analysis,” *IEEE Transactions on Image Processing*, 2019.

[20] A. Wong and S. Soatto, “Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5644–5653.

[21] R. Mahjourian, M. Wicke, and A. Angelova, “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*, 2018, pp. 5667–5675.

[22] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, “Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding,” in *European Conference on Computer Vision*. Springer, 2018, pp. 691–709.

[23] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[24] A. Wong, S. Cicek, and S. Soatto, “Targeted adversarial perturbations for monocular depth prediction,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[25] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, “Deeppruner: Learning efficient stereo matching via differentiable patchmatch,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4384–4393.

[26] A. Shaked and L. Wolf, “Improved stereo matching with constant highway networks and reflective confidence learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4641–4650.

[27] A. Seki and M. Pollefeys, “Patch based confidence prediction for dense disparity map.” in *BMVC*, vol. 2, no. 3, 2016, p. 4.

[28] A. Spyropoulos, N. Komodakis, and P. Mordohai, “Learning to detect ground control points for improving the accuracy of stereo matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1621–1628.

[29] M. Klodt and A. Vedaldi, “Supervising the new with the old: learning sfm from sfm,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 698–713.

[30] X. Hu and P. Mordohai, “A quantitative evaluation of confidence measures for stereo vision,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2121–2133, 2012.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[32] X. Fei, A. Wong, and S. Soatto, “Geo-supervised visual depth prediction,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1661–1668, 2019.