# Detail-Preserving Self-Supervised Monocular Depth with Self-Supervised Structural Sharpening

Juan Luis Gonzalez Bello juanluisgb@kaist.ac.kr
Jaeho Moon jaeho.moon@kaist.ac.kr
Munchurl Kim mkimee@kaist.ac.kr

Korea Advanced Institute of Science and Technology

## Abstract

*We propose to further close the gap between self-supervised and fully-supervised methods for the single view depth estimation (SVDE) task in terms of the levels of detail and sharpness in the estimated depth maps. Detailed SVDE is challenging as even fully-supervised methods struggle to obtain detail-preserving depth estimates. While recent works have proposed exploiting semantic masks to improve the structural information in the estimated depth maps, our proposed method yields detail-preserving depth estimates from a single forward pass without increasing the computational cost or requiring additional data. We achieve this by exploiting a missing component in SVDE, Self-Supervised Structural Sharpening, referred to as $S^4$. $S^4$ is a mechanism that encourages a similar level of detail between the RGB input and the depth/disparity output. To this extent, we propose a novel DispNet-$S^4$ network for detail-preserving SVDE. Our network exploits un-blurring and un-noising tasks of clean input images for learning $S^4$ without the need for either additional data (e.g., segmentation masks, matting maps, etc.) or advanced network blocks (attention, transformers, etc.). The recovered structural details in the un-blurring and un-noising operations are transferred to the estimated depth maps via adaptive convolutions to yield structurally sharpened depths that are selectively used for self-supervision. We provide extensive experimental results and ablation studies that show our proposed DispNet-$S^4$ network can yield fine details in the depth maps while achieving quantitative metrics comparable to the state-of-the-art for the challenging KITTI dataset.*

## 1. Introduction

Predicting the world geometry or estimating the depths from images is a fundamental problem in computer vision, robotics, and computational imaging. The depth estimation (DE) task is essential for downstream problems, such as robotic navigation and grasping, de-focus blur/de-blur, novel view synthesis [4, 26, 47, 50], de-hazing [28, 29], and semantic segmentation [2, 18, 34, 43].

The advance of deep learning has shown its supremacy over classical approaches that rely on sparse correspondences and hard-coded assumptions [6, 27, 48] for both multi-view and single view depth estimation (SVDE) tasks. In particular, SVDE is a more challenging and ill-posed problem than its multi-view counterpart, as it requires a global and local understanding of the scene's low and high-level depth cues. Furthermore, it is even more challenging to learn SVDE in a self-supervised manner, that is, without the hard-to-obtain depth ground truths (GT).

Apart from the well-known issues that the self-supervised monocular DE methods suffer from the artifacts and inaccuracies in highly homogeneous and/or reflective regions, occlusions, and independently moving objects, we identify, in this paper, that the self-supervised DE models lack an effective way to enforce structural similarity between the input RGB image and the output inverse depth (also known as disparity). This lack of structural similarity enforcement, combined with the weak supervision from the adjacent frames' photometric reconstructions, leads the SVDE networks to generate depth estimates with much fewer details and wider borders than the input RGB image, as depicted in Fig. 1. The inaccurate borders of the rendered depths harm downstream tasks that require high structural details, such as Bokeh effects and augmented reality.

While previous works attempt to preserve such high structural details by incorporating segmentation or matting maps [16, 53], we realize that such structural information is already given in the input image and that low-level vision pre-text tasks can be exploited to obtain a higher level of details in the estimated depths. Our main contribution in this paper is in **Self-Supervised Structural Sharpening** ($S^4$). $S^4$ is a combination of a novel learning pipeline, network architecture, and loss functions to enforce detail-preserving depth estimation. $S^4$ closes the gap of details between the network's inputs and outputs by predicting *un*-blurring and *un*-noising per-pixel adaptive kernels that re-

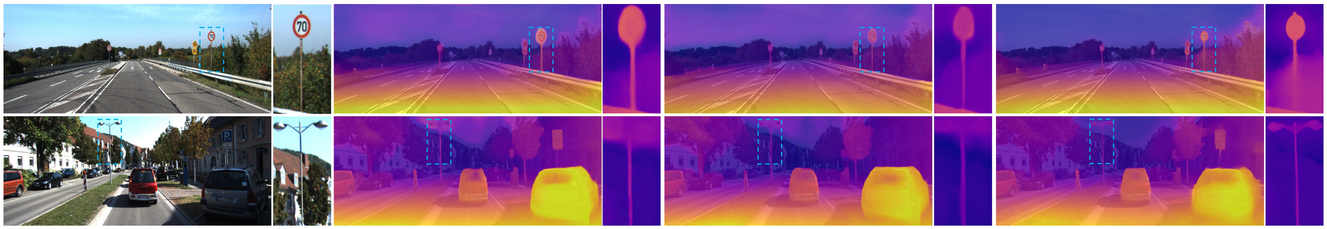

Figure 1. Our model learns to generate detail-preserving disparity maps by training with Self-Supervised Structural Sharpening ($S^4$).

store a blindly blurred and noisy version of the input image. The blindly blurred and noisy version of the input image is used as a guide for the $S^4$ to learn the structural details during training. Note that in vanilla de-blurring and de-noising tasks, the blurred and noisy image is the input to the network, and the restored image is the output. In contrast, in our *un*-blurring and *un*-noising, the input to the network is the clean RGB image, and the outputs are the locally adaptive kernels. The resulting locally adaptive convolutions are then applied to the network's estimated disparity map to obtain a structurally sharpened disparity, which shares the same structural details as the restored blur and noisy image. The structurally sharpened disparity is back again utilized to self-supervise the network's estimated disparity in a *recursively interactive* manner, thus enforcing the estimated disparity map to reach a level of high structural fidelity.

$S^4$ has the properties of (i) significantly improving the quality of the estimated depth maps (as depicted in Fig. 1), (ii) removing the burden of estimating detailed structures from the network, and (iii) achieves comparable to SOTA performance on the challenging KITTI dataset.

## 2. Related Work

### 2.1. Fully-supervised learning of SVDE

The early work [3] on learning SVDE was proposed to directly regress depth with a multi-scale deep network that utilized a scale-invariant error to alleviate global scale ambiguity. On the other hand, the work of Fu *et al*. [5] was proposed to solve SVDE as an ordinal regression task by discretizing the output and the GT values for training, yielding considerable performance gains. In the work of BTS [25], instead of discretizing or regressing depth values, Lee *et al*. proposed to predict local planar guidance that locally describes the 3D geometry of the scene in a low to full resolution fashion. Ranftl *et al*. [36] proposed a strategy for multi-dataset training by optimizing an objective function that is invariant to range and scale in depth values for each dataset. The more recent work of Miangoleh *et al*. [32] attempted to boost the performance of [36] by aggregating additional forward passes of low and high resolution to enhance the network's receptive field while preserving the overall structure details. Xian *et al*. [49] proposed to enhance the details in the estimated depth maps with a structure ranking loss that is guided by estimated instance segmentation masks and image edges. More recent works have also explored the joint learning of SVDE, and depth completion as in [14], while others have explored incorporating vision transformers for dense prediction [35].

### 2.2. Self-supervised learning of SVDE

Self-supervised learning of SVDE relies on photometric reconstruction losses that compare 3D projected training images, guided by the estimated depths, with their corresponding target images. Self-supervised SVDE can be divided into learning from synchronized images (e.g. stereo images) and learning from unsynchronized images (e.g. monocular videos).

For the stereo cases, the works of Watson *et al*. [45] and Tosi *et al*. [41] proposed to improve the earlier work of [9] by incorporating and distilling the result of classical stereo disparity estimation techniques, such as SGM [19], to provide proxy labels during training. Gonzalez and Kim [10] proposed an SVDE method with an exponential disparity quantization and a multi-view occlusion mask computation technique that exploited the geometric properties of disparity probability volumes. Zhu *et al*. [53] proposed edge-edge consistency between semantic segmentation maps and depth maps via a morphing algorithm. Saeedan and Roth [37] also exploited additional training information in the form of panoptic segmentation masks in their guided alignment and smoothness losses. On the other hand, Gonzalez and Kim [11] proposed a distilled matting Laplacian loss for enhancing depth maps on the objects' borders and the use of pixel positional information for better learning from randomly resized and cropped patches.

For the monocular video cases, Godard *et al*. proposed auto-masking in their Monodepth2 [8] to improve the explainability masks in the earlier work of [52]. In contrast, the work in [12] proposed a method for masking moving objects in the scene by measuring and thresholding the variance of the network's output, achieving performance gains in the self-supervised SVDE task. In [39], a learnable feature-metric loss is proposed instead of using L1 or

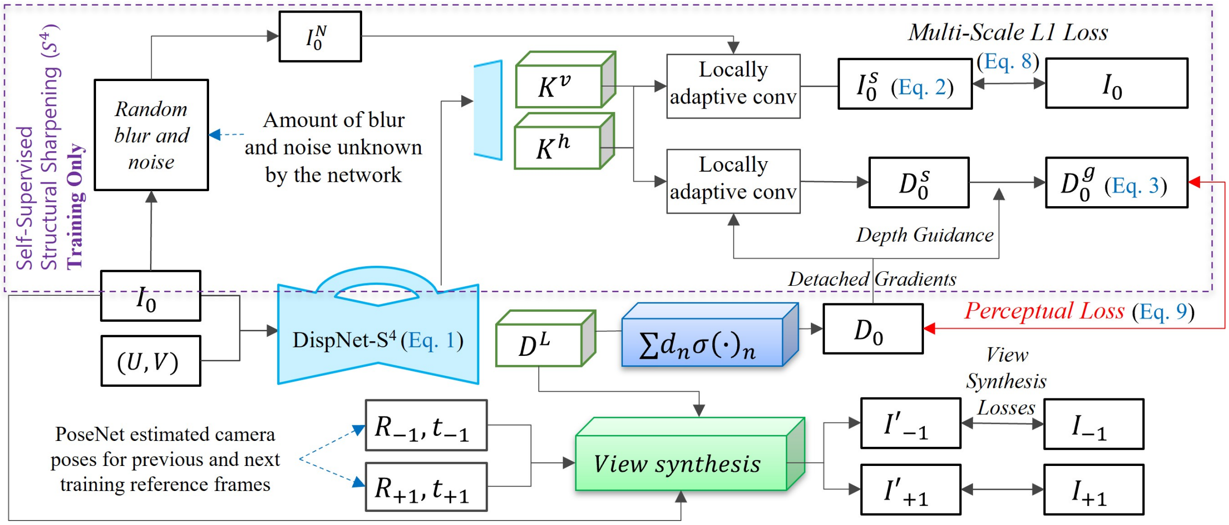

Figure 2. Proposed learning pipeline with our main contribution, the Self-Supervised Structural Sharpening ($S^4$), highlighted.

SSIM [44] loss. Guizilini *et al*., in [15], proposed the PackNet which contains 3D packing and unpacking blocks that exploit sub-pixel [38] and 3D convolutions for end-to-end structural preservation of details, yielding the SOTA results for self-supervised SVDE from videos.

In their later work, Guizilini *et al*. [16] exploited semantic segmentation into the PackNet decoder part as well as a two-stage learning strategy for better handling of moving objects in the scene. The work of Klinger *et al*. [24] also proposed to handle moving objects with semantic guidance by learning semantic segmentation and SVDE simultaneously, assuming that certain object classes move more often than others (e.g. cars, bikers, etc.). Jung *et al*. [22] also proposed using semantics information in their triplet loss to refine depth representations according to implicit semantic guidance and a cross-task attention module for guiding depth features to be more semantically consistent. The recent work of Lyu *et al*. [31] proposed to improve Monodepth2 [8] by implementing dense connections between the encoder and the decoder and fusing them via squeeze and excitation blocks [20], improving the level of details and quantitative metrics considerably. Watson *et al*. [46] explored learning depth estimation from multiple frames as input by utilizing a Monodepth2 [8] as a teacher network to identify unreliable pixels induced by static sequences or moving objects.

## 3. Method

We propose a simple yet effective addition to self-supervised SVDE pipelines, which exploits the untapped low-level vision tasks of *un*-blurring and *un*-noising, referred to as Self-Supervised Structural Sharpening, or $S^4$. The overview of our proposed pipeline with $S^4$ is depicted in Fig. 2. We refer to the DispNet that utilizes $S^4$ as DispNet-$S^4$. The DispNet-$S^4$, with a convolutional autoencoder backbone similar to that of the works of [8, 11], takes as input a single RGB image $\mathbf{I}_0$ along with its normalized pixels' coordinates information $(\mathbf{U}, \mathbf{V})$ (as in [11]). DispNet-$S^4$ outputs a probability disparity volume (logits) $\mathbf{D}^L$ (as in [10]) and per-pixel adaptive kernels which are separated into vertical and horizontal components, $\mathbf{K}^v$ and $\mathbf{K}^h$ respectively. The two separable kernels are estimated via a separate branch in the last decoder stage, as depicted in Fig. 2. $\mathbf{D}^L$, $\mathbf{K}^v$, and $\mathbf{K}^v$ are then described by

$$\mathbf{D}^L, \mathbf{K}^h, \mathbf{K}^v = DispNet\text{-}S^4(\mathbf{I}_0, (\mathbf{U}, \mathbf{V})). \tag{1}$$

The disparity logits $\mathbf{D}^L \in \mathbb{R}^{n \times H \times W}$ (for an input image of resolution $H \times W$) are soft-maxed and channel-wise dot-produced with their corresponding $n$ disparity quantization levels $d_n$, yielding the final disparity estimate $\mathbf{D}_0 = \boldsymbol{d} \odot \mathbf{D}^L$. As shown in the bottom part of Fig. 2, $\mathbf{D}^L$ is also utilized for forward warping to project $I_0$ onto the next and previous reference training frames $\mathbf{I}_{+1}$ and $\mathbf{I}_{-1}$, respectively. Also, for this operation, the camera poses are estimated by a PoseNet following the previous works of [8, 12, 52]. The resulting novel views $\mathbf{I}'_{+1}$ and $\mathbf{I}'_{-1}$ are then compared against their corresponding GTs with an occlusion and moving object-aware photometric reconstruction loss, following the previous literature [12].

Although originally devised for frame interpolation [33], adaptive separable convolution kernels $\mathbf{K}^v$ and $\mathbf{K}^h$ can also be utilized to approximate an $n_v \times n_h$ adaptive filter. $n_h$ and $n_v$ are the number of kernel values in $\mathbf{K}^v$ and $\mathbf{K}^h$, respectively. In our $S^4$, $\mathbf{K}^v$ and $\mathbf{K}^h$ filter a blindly noised and blurred version of $\mathbf{I}_0$, denoted by $\mathbf{I}_0^N$, yielding the structurally sharpened $\mathbf{I}_0^s$ image. $\mathbf{I}_0^s$ is compared against the input image with a multi-scale L1 loss. We refer to this pre-text task as *un*-blurring and *un*-noising. Note that DispNet-$S^4$ cannot account for the amount of blur or noise as $\mathbf{I}_0^N$ is not an input, which differs from the standard de-noising and deblurring task which takes as input blur and noisy images and outputs clean images. See Section 3.1 for more details on $S^4$ and learning *un*-blurring and *un*-noising.

Our novel pipeline with $S^4$ tackles the lack of detail preservation in self-supervision by transferring the structural details from the restored image $\mathbf{I}_0^s$ into the estimated disparity map $\mathbf{D}_0$ by applying the same locally adaptive kernels, yielding the structurally sharpened $\mathbf{D}_0^s$. Note that this mechanism serves as the bridge for the interaction between the RGB domain and the depth domain via the locally learned adaptive convolution filters. In addition, we guide the sharpening operation with the initial depth $\mathbf{D}_0$, which helps alleviate non-structural artifacts that arise from the high color guidance in the *un*-blurring and *un*-noising task. This yields the depth-guided sharpened disparity map $\mathbf{D}_0^g$. See Section 3.2 for details on depth guided $S^4$.

Finally, $\mathbf{D}_0^g$ is back again to supervise the disparity estimate $\mathbf{D}_0$ via either multi-scale L2 or perceptual losses. These losses also encourage the *un*-blurring and *un*-noising operations to maintain the geometrical details in the image. More details on loss functions are available in Section 3.3.

## 3.1. Self-Supervised Structural Sharpening ($S^4$)

The view synthesis losses in Fig. 2 provide the main source of 3D self-supervision, as the network has to learn the correct depth probabilities in $\mathbf{D}^L$ to properly perform forward warping for all images in the training dataset. However, these losses are unable to enforce the SVDE network learning the high level of structural details, such as the ones needed to properly represent the geometries of thin structures and accurate borders of the objects in the scene. This is because 'thicker' depth maps that leak into the background better warp objects with a large variety of dimensions providing an 'easy' way to minimize the loss functions.

Enforcing the structural similarity between the RGB input and the disparity output is not a straightforward task. Disparity (inverse depth) maps are invariant to shadows, reflections, and color changes in the scene, which are abundant in RGB images. To bring rich details of input RGB images into the disparity domain, we propose, for the first time, blind *un*-blurring and *un*-noising as a pre-text task via locally adaptive convolutions. By solving this task, we can transfer the structural details from the RGB image into the depth output, creating a bridge for self-supervised structural sharpening.

### 3.1.1 Learning blind *un*-blurring and *un*-noising

Similar to how the prefix *de-* tends to indicate action and *un-* connotes a passive status, we propose *un*-blurring and *un*-noising, where we learn how to blindly restore blur and noisy image versions of the already clean and un-distorted input image as shown in Fig. 3-(a). In this setting, the network has to predict kernel values that properly filter the input image, regardless of the window size and standard deviation of the applied Gaussian blur and noise. We hypothesize that learning the blind *un*-blurring and *un*-noising enforces the network to learn a rudimentary color-guided segmentation, as suggested by the structurally sharp but smooth restored images $I_0^s$ in Fig. 3-(a). The locally adaptive and separable filtering operation can be expressed as

$$\mathbf{I}_0^s(u,v) = \sum_{j=1}^{n_v} \boldsymbol{k}^{v,s}_{u,v,j} \sum_{i=1}^{n_h} \boldsymbol{k}^{h,s}_{u,v,i} \mathbf{I}_0^N(u+i-\tfrac{n_h}{2}, v+j-\tfrac{n_v}{2}) \tag{2}$$

where $\mathbf{I}_0^s(u,v)$ denotes a restored pixel at location $(u,v)$ using point-wise separable vertical and horizontal kernels $\boldsymbol{k}^v_{u,v}$ and $\boldsymbol{k}^h_{u,v}$. The adaptive separable kernels are pre-processed by $\boldsymbol{k}^{v,s}_{u,v} = \sigma(\boldsymbol{k}^v_{u,v})$ and $\boldsymbol{k}^{h,s}_{u,v} = \sigma(\boldsymbol{k}^h_{u,v})$ where $\sigma(\cdot)$ denotes the soft-max operation, that is, the resulting restored pixel color is a combination of the colors in the local $n_v \times n_h$ window. In Eq. 2, the inner summation describes the horizontal filtering, and the outer summation performs the vertical filtering.

In particular, *un*-blurring forces the network to learn sharpening in the color edges, while *un*-noising guides it to learn smoothing in the homogeneous regions. Together, they guide the network to perform structural sharpening of the blindly blurred and noisy image $\mathbf{I}_0^N$ into the clean prediction $\mathbf{I}_0^s$. Surprisingly, this is a very simple task for the network to learn, as it already knows all the structural details from the un-distorted input image $\mathbf{I}_0$. We have observed that the loss for this pre-text task reaches values close to its optimal point in just a couple of epochs. We apply a random Gaussian blur window size from $K = 1 \times 1$ (or identity) to $K = 31 \times 31$, where the standard deviation is set to $\sigma = k/3$. For the random noise, we utilized speckle noise, which randomly scales RGB values with a probability of $10\%$ but any type of noise should suffice.

### 3.1.2 Structural sharpening of depth

The recovered structures in the clean prediction $\mathbf{I}_0^s$ can be *transferred* to the network's estimated disparity map $\mathbf{D}_0$ by applying the same local adaptive filtering for $\mathbf{D}_0$ as done for $\mathbf{I}_0^s$. The result is the structurally sharpened disparity $\mathbf{D}_0^s$, which shares a similar level of detail as the restored image $\mathbf{I}_0^s$, as seen in the top-right region of Fig. 3-(b).

We can take advantage of $\mathbf{D}_0^s$ by using it as a "teacher" for the network's output $\mathbf{D}_0$, as depicted by the perceptual loss in Fig. 2. We noted that the perceptual loss [21] was more effective than other losses such as L1 or L2, as it compares not only the individual pixel disparity values but also

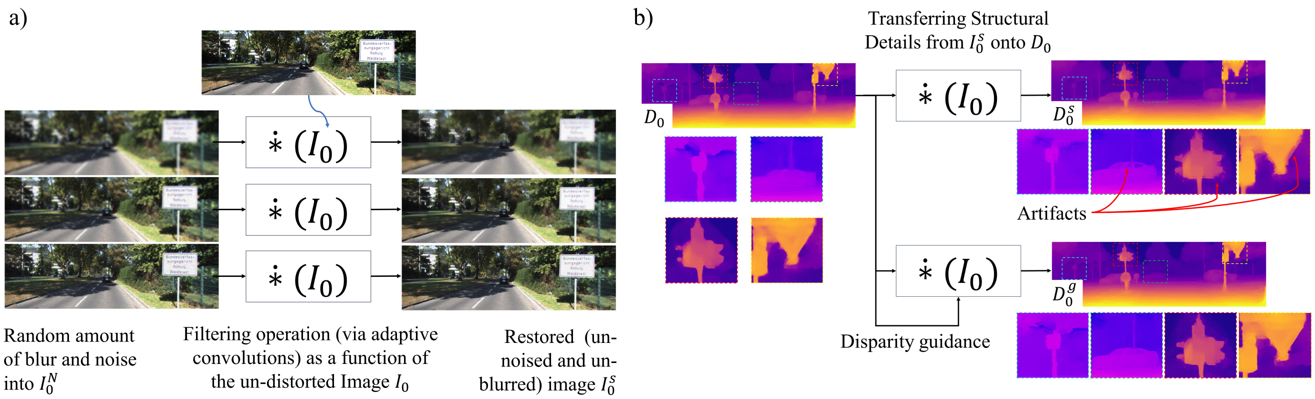
a)
b)
$\dot{*}\,(I_0)$
$\dot{*}\,(I_0)$
$\dot{*}\,(I_0)$
Random amount of blur and noise into $I_0^N$
Filtering operation (via adaptive convolutions) as a function of the un-distorted Image $I_0$
Restored (un-noised and un-blurred) image $I_0^s$
Transferring Structural Details from $I_0^s$ onto $D_0$
$D_0$
$\dot{*}\,(I_0)$
$D_0^s$
Artifacts
$\dot{*}\,(I_0)$
$D_0^g$
Disparity guidance

Figure 3. *Un*-blurring and *un*-noising in our $S^4$. a) Restoration of blindly blurred and noisy images. b) Sharpening of the estimated disparity $\mathbf{D}_0$. The guided $\mathbf{D}_0^g$ is structurally sharper than $\mathbf{D}_0$ with considerably fewer artifacts than $\mathbf{D}_0^s$.

their structures and relationships with their nearby pixels as given by the VGG's [40] receptive field.

### 3.2. Depth guided $S^4$

Disparity or depth information can also be utilized to guide the adaptive filtering operation by lowering the value of the locally adaptive filter weights for the pixels that are much closer or farther than the target center pixel. While $\mathbf{D}_0^s$ is obtained by adaptively filtering the $\mathbf{D}_0$ according to Eq. 2, a guided $\mathbf{D}_0^g$ can be described by re-writing Eq. 2 as:

$$\mathbf{D}_0^g(u,v)=\sum_{j=1}^{n_v}\boldsymbol{k}_{u,v,j}^{v,g,s}\sum_{i=1}^{n_h}\boldsymbol{k}_{u,v,i}^{h,g,s}\mathbf{D}_0(u+i-\tfrac{n_h}{2},v+j-\tfrac{n_v}{2}) \tag{3}$$

where $\boldsymbol{k}_{u,v,i}^{h,g,s}$ and $\boldsymbol{k}_{u,v,j}^{v,g,s}$ denote the horizontal $i^{th}$ and vertical $j^{th}$ disparity-guided kernel elements for a pixel at $(u,v)$, respectively. We define $\boldsymbol{k}_{u,v}^{h,g,s}=\sigma(\boldsymbol{k}_{u,v}^{h}+\boldsymbol{g}_h(u,v))$, where the horizontal guidance, $\boldsymbol{g}_h(u,v))$ is given by

$$\boldsymbol{g}_h(u,v)=\alpha-\beta\left(\frac{\boldsymbol{d}_{herr}}{max(\boldsymbol{d}_{herr})+\epsilon}\right)^2, \tag{4}$$

where $\boldsymbol{d}_{herr}$ are the errors between the target center pixel at $(u,v)$ and its local horizontal neighbors, as given by

$$\boldsymbol{d}_{herr}(u,v)=\left|\left|\mathbf{D}_0(u,v)-\left[\mathbf{D}_0(u+x,v)\right]_{x=-\frac{n_h}{2}}^{x=\frac{n_h}{2}}\right|\right|, \tag{5}$$

$\alpha$ and $\beta$ are the coefficients of the inverted parabola which is a function of the normalized errors, and $\epsilon=1\times10^{-5}$ provides numerical stability. $\boldsymbol{k}_{u,v,j}^{v,g,s}$ is obtained in a similar manner following Eqs. 4 and 5 for the vertical kernel elements. We set $\alpha=1$ and $\beta=4$ to provide positive guidance when the normalized error is smaller than $50\%$ and negative guidance when the normalized error is larger than $50\%$. Depth-guided $S^4$ helps in obtaining detailed disparity maps with considerably fewer artifacts, as shown in the bottom-right region of Fig. 3-(b). Note that, as shown in Fig. 2, $\mathbf{D}_0^g$ is only computed during training and used as a pseudo-GT to infuse a greater level of detail into $\mathbf{D}_0$.

### 3.3. Loss Functions

The total loss function $l$ for the self-supervision of our DispNet-$S^4$ is a combination of view synthesis loss which provides the main 3D self-supervision $l_{syn}^{-1}+l_{syn}^{+1}$, edge-preserving disparity smoothness loss $l_{ds}$ (weighed by $\alpha_{ds}=0.1$ as in [9, 10]), and the proposed $S^4$ loss $l_{S^4}$, as given by

$$l=l_{syn}^{-1}+l_{syn}^{+1}+\alpha_{ds}l_{ds}+l_{S^4}. \tag{6}$$

#### 3.3.1 The $S^4$ loss

The proposed novel $S^4$ loss ($l_{S^4}$) consists of the *un*-blur and *un*-noise loss $l_{un}$ and the depth detail loss $l_d$ as described in Eq. 7. Using the *un*-blur and *un*-noise loss, our DispNet-$S^4$ learns the locally adaptive point-wise kernels from a clean input RGB image to restore the blindly blurred and noisy image $\mathbf{I}_0^N$. The depth detail loss provides self-supervision via the perceptual loss between the depth-guided sharpened disparity $\mathbf{D}_0^g$ and the DispNet-$S^4$ estimated disparity $\mathbf{D}_0$. Note that, as depicted in Fig. 2, the gradients of the $\mathbf{D}_0$ tensor are disabled when computing $\mathbf{D}_0^g$, such that $l_d$ focuses more on matching the structural details between $\mathbf{D}_0$ and $\mathbf{D}_0^g$. The $S^4$ loss is defined as

$$l_{S^4}=\alpha_{un}l_{un}+\alpha_d l_d, \tag{7}$$

where each variable is detailed in the next.

***Un*-blur and *un*-noise loss** ($l_{un}$) guides the network to clean the blindly blurred and noised image $\mathbf{I}_0^N$ via a weighted multi-scale L1 loss between the restored image $\mathbf{I}_0^s$ and the original input image $\mathbf{I}_0$, as given by

$$l_{un}=\sum_{k=0}^{3}||\mathbf{D}_{hf}^k\odot(f_k(\mathbf{I}_0^s)-f_k(\mathbf{I}_0))||_1, \tag{8}$$

where $\odot$ denotes the element-wise product in Eq. 8, while $f_k(\cdot)$ represents a $2^k$ down-sampling operation. At each scale $k$, $l_{un}$ is weighted by the normalized high-frequency component of $\mathbf{D}_0$, denoted by $\mathbf{D}_{hf}^k = f_k(\mathbf{D}_{hf})$. The weighting of $l_{un}$ by $\mathbf{D}_{hf}$ loosely guides the network to restore structural details more than other non-structural details (e.g. textures, shadows, etc.). We empirically set $\alpha_{un} = 1$.

**Depth detail loss ($l_d$).** This term enforces structural similarities between the output and sharpened disparities $\mathbf{D}_0$ and $\mathbf{D}_0^g$ via a perceptual loss [21], as given by

$$l_d = \sum_{l=0}^{3} ||\mathbf{D}_n^l \odot (\phi^l(\mathbf{D}_0) - \phi^l(\mathbf{D}_0^g))||_2^2, \tag{9}$$

where $\phi^l(\cdot)$ denote the $l^{th}$ maxpool layer of a pre-trained VGG19 [40] in the perceptual loss. This loss is weighted by the mean-normalized resized disparity $\mathbf{D}_n^l = g_l(\mathbf{D}_n)$ which is given by

$$\mathbf{D}_n(u,v) = \begin{cases} 1 & if\ \mathbf{D}_0(u,v)/\bar{D}_0 > 1 \\ (\mathbf{D}_0(u,v)/\bar{D}_0)^3 & o.w. \end{cases} \tag{10}$$

where $\bar{D}$ is the mean disparity. Weighting the loss $l_d$ by $\mathbf{D}_n$ helps alleviate depth artifacts in the structurally sharpened disparities that are caused by the excessive color guidance. While these artifacts can appear in any region, they are more noticeable in faraway regions. This is because the photometric reconstruction losses, which provide a weak 3D self-supervision, are much smaller for the distant objects, which can be estimated much closer or further without affecting the resulting synthesized novel views much. We observed that using the estimated depth information to guide the distortion operations (noise and blur) in $\mathbf{I}_0^N$ yielded similar results as weighting $l_d$ by $\mathbf{D}_n$. However, for simplicity, we only weigh $l_d$ by $\mathbf{D}_n$ in this paper.

## 4. Experiments and Results

We implemented our network and training methods with PyTorch and trained on an NVIDIA A100 GPU with a batch size of 8 by the Adam [23] optimizer (with $\beta_1$ and $\beta_2$ set to 0.9 and 0.999, respectively) for 105 epochs on the KITTI Eigen train split and on CityScapes. The initial learning rates were respectively set to $1 \times 10^{-4}$ and $5 \times 10^{-5}$ for KITTI [7] and CityScapes [1], and halved at 55, 75, and 95 epochs. We used 31 channels in $\mathbf{D}^L$ and 15 kernel elements for $\mathbf{K}^h$ and $\mathbf{K}^h$ each. For a fair comparison with the previous works [8, 12, 15, 22, 31], we adopted random resizing from a factor of 0.5 to 1.5, followed by 640×192 random cropping, random horizontal flip, random gamma, color, and brightness shifts.

**The KITTI [7] dataset.** We utilized the Eigen split [3] from the KITTI dataset for training and testing. The Eigen split contains 22,600 train images and 697 test images. Furthermore, we adopted the improved version of this test split [42] which contains 652 images with denser depth ground truths that are obtained by selectively aggregating LiDAR points from 5 consecutive frames.

**The CityScapes [1] dataset.** To test the generalization capability of $S^4$ we also trained the Dispnet-S$^4$ on the CityScapes dataset. The CityScapes dataset contains about 3K images surrounded by 29 frames each. The whole training sequences consist of 80K images. To realize large enough amounts of motion between frames, we halved the frame rate by randomly skipping odd or even frames.

### 4.1. Results on KITTI

We present our quantitative results in Table 1 and our qualitative results in Fig. 4. As shown in Table 1, we achieve the best performance in all metrics for the improved Kitti Eigen test split [42], and the best RMSE metric on the original split [3]. However, it should be noted that the top performers in the improved split such as PackNet [15], AQUANet [12], and our DispNet-$S^4$ have shown slightly inferior performance in the original split. This is due to up to 5× sparser depth GT in the original split, which benefits other methods that yield blurrier results.

In addition, we show the generalization ability of our proposed method by incorporating $S^4$ into the Monodepth2 [8] and PLADENet [11], denoted by Monodepth2-$S^4$ and PLADENet-$S^4$, respectively in Table 1. The consistent performance improvement of Monodepth2-$S^4$ vs Monodepth2 shows that $S^4$ can also benefit methods that learn from monocular videos with auto-masking and backward-warping-based loss functions [8]. On the other hand, the higher performance of PLADENet-$S^4$ over PLADENet shows that our method can also benefit methods that learn SVDE from synchronized stereo images.

Qualitatively, our method also outperforms the recent works [15, 22, 25, 31] by producing very detailed depths in all image regions, as depicted in Fig. 4. In particular, our DispNet-$S^4$ is the only model that can represent the geometrical details in the threes in rows 1, 3, and 4. Our network with $S^4$ also gets the best contours for the traffic signs and lights in rows 2, 5, and 6, while the other methods yield rather over-smooth depth maps. Furthermore, our method generates depth maps with the least amount of depth artifacts in the sky regions.

### 4.2. Ablation Studies

We ablate the effects of our novel training strategy with $S^4$ on KITTI [7] and present the results in Table 2 and Fig. 5. The 'Baseline' DispNet, does not incorporate our $S^4$ training strategy and loss functions. Instead, the baseline is trained for self-supervised SVDE from videos following [12] without their adaptive quantization. As can be seen in

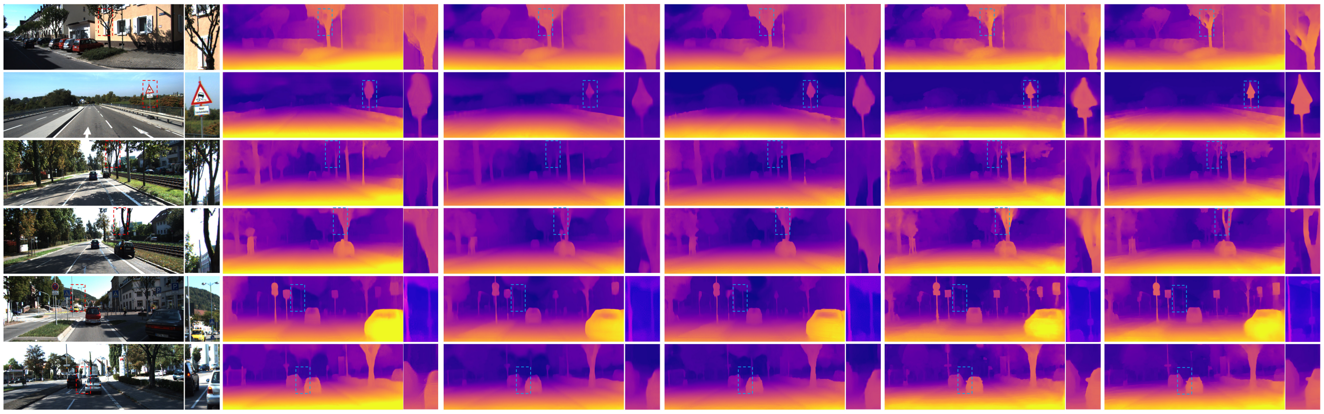

Figure 4. Qualitative comparisons on KITTI [7] among self-supervised methods.

| | Ref | Method | Sup. | Par(M) | abs rel↓ | sq rel↓ | rmse↓ | $\text{rmse}_{log}$ ↓ | $\delta^1$ ↑ | $\delta^2$ ↑ | $\delta^3$ ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Original Eigen Test Split [7] | [30] | Luo *et al.* | D+S | - | **0.094** | **0.626** | 4.252 | 0.177 | **0.891** | 0.965 | 0.984 |
| | [17] | Gur *et al.* | DoF | - | 0.110 | 0.666 | **4.186** | **0.168** | 0.880 | **0.966** | **0.988** |
| | [8] | Monodepth2 | V | 14 | 0.115 | 0.882 | 4.701 | 0.190 | 0.879 | 0.961 | 0.982 |
| | Our | Monodepth2-$S^4$ | V | 14 | *0.112* | *0.876* | *4.754* | *0.187* | *0.882* | *0.962* | *0.982* |
| | [31] | HR-Depth | V | 14 | 0.111 | 0.833 | 4.673 | 0.187 | 0.882 | 0.961 | 0.982 |
| | [22] | FSRE-Depth$_{R18}$ | V+Se | 25 | 0.105 | 0.722 | 4.547 | 0.182 | 0.886 | 0.964 | 0.984 |
| | [15] | PackNet | V | 120 | 0.107 | 0.802 | 4.538 | 0.186 | 0.889 | 0.962 | 0.981 |
| | [22] | FSRE-Depth$_{R50}$ | V+Se | 25 | 0.102 | 0.675 | 4.393 | 0.178 | 0.893 | **0.966** | **0.984** |
| | [16] | Guizilini *et al.* * | V+Se | 140 | **0.100** | 0.761 | 4.270 | **0.175** | **0.902** | 0.965 | 0.982 |
| | [12] | *AQUANet* | V | 14 | 0.115 | **0.656** | 4.251 | 0.186 | 0.875 | 0.959 | 0.983 |
| | our | **DispNet-$S^4$** | V | 14 | 0.112 | 0.676 | **4.206** | 0.181 | 0.881 | 0.963 | **0.984** |
| Improved Test Split [42] | [51] | Yin *et al.* | D | 113 | 0.072 | - | 3.258 | 0.117 | 0.938 | 0.990 | 0.998 |
| | [5] | DORN | D | 51 | 0.072 | 0.307 | 2.727 | 0.120 | 0.932 | 0.984 | 0.995 |
| | [8] | Monodepth2 | V+S | 14 | 0.087 | 0.479 | 3.595 | 0.131 | 0.916 | 0.984 | 0.996 |
| | [45] | DepthHints | $\text{S}_{\text{SGM}}$ | 35 | 0.074 | 0.364 | 3.202 | 0.114 | 0.936 | 0.989 | 0.997 |
| | [53] | Edge-of-Depth | S | - | 0.076 | 0.348 | 3.117 | 0.113 | 0.938 | 0.990 | 0.997 |
| | [11] | PLADENet | S | 15 | **0.068** | 0.291 | 2.974 | **0.107** | 0.942 | **0.991** | **0.998** |
| | our | PLADENet-$S^4$ | S | 15 | **0.068** | **0.289** | **2.914** | **0.107** | **0.943** | **0.991** | **0.998** |
| | [8] | Monodepth2 | V | 14 | 0.092 | 0.536 | 3.749 | 0.135 | 0.916 | 0.984 | 0.995 |
| | [31] | HR-Depth | V | 14 | 0.087 | 0.507 | 3.787 | 0.132 | 0.919 | 0.983 | 0.995 |
| | [22] | FSRE-Depth$_{R18}$ | V+Se | 25 | 0.084 | 0.436 | 3.740 | 0.129 | 0.919 | 0.985 | 0.996 |
| | [15] | PackNet | V | 120 | 0.078 | 0.420 | 3.485 | 0.121 | 0.931 | 0.986 | 0.996 |
| | [12] | *AQUANet* | V | 14 | 0.077 | 0.324 | 3.032 | 0.115 | 0.938 | 0.989 | 0.997 |
| | our | **DispNet-$S^4$** | V | 14 | **0.075** | **0.316** | **3.020** | **0.112** | **0.940** | **0.990** | **0.998** |

Table 1. Comparison to existing SVDE methods on the KITTI Eigen Split [3]. DoF: supervised by depth of field. D: depth-supervised. V, V+Se, S, and $\text{S}_{\text{SGM}}$: Self-supervised from video, video+semantics, stereo, and stereo+SGM. V models use median scaling. Error metrics (defined in [3]) are ↓ the lower the better, and accuracy metrics are ↑ the higher the better. * pre-trained on [1].

the first column of Fig. 5, the baseline network yields noisy depth maps with no clear object borders. Enabling $l_{un}$ while keeping $l_d$ disabled (no self-supervised structural sharpening) already brings marginal quantitative improvements, as denoted by 'Multi-task learning' in Table 2, but provides no structural sharpening benefits in the estimated depth maps. For the sake of completeness, we used a 'Multi-scale L2' loss instead of perceptual loss for our $l_d$ and obtained more considerable quantitative and qualitative gains with respect to the baseline. Moreover, the usage of perceptual loss be-

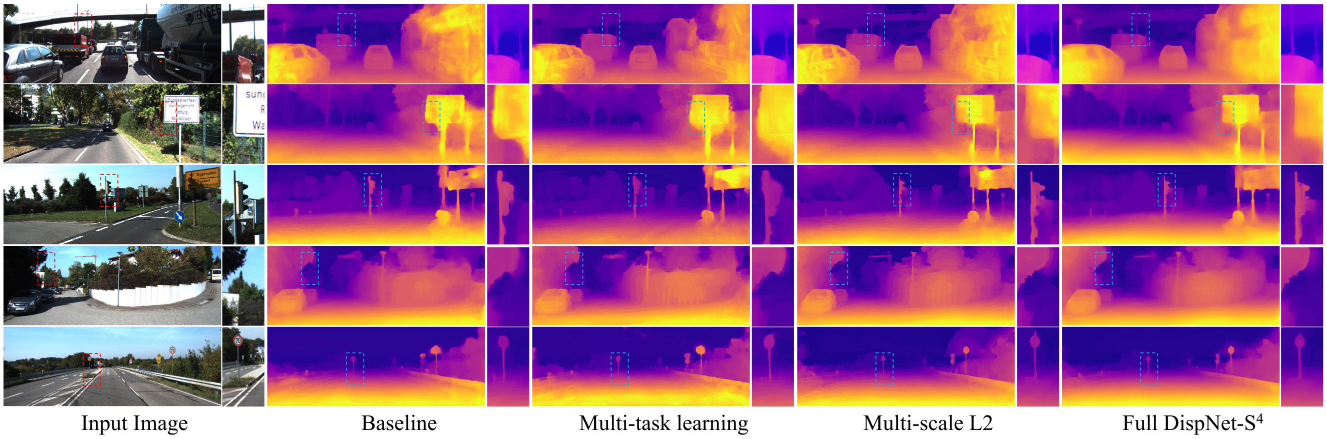

Figure 5. Ablation studies on our $S^4$ loss.

| Ablation Study | abs rel↓ | sq rel↓ | rmse↓ | rmse$_{log}$ ↓ | $\delta^1$ ↑ | $\delta^2$ ↑ | $\delta^3$ ↑ |
|---|---|---|---|---|---|---|---|
| Baseline $(\alpha_{un} = 0, \alpha_d = 0)$ | 0.079 | 0.347 | 3.105 | 0.119 | 0.934 | 0.987 | 0.997 |
| Multi-task learning $(\alpha_{un} = 1, \alpha_d = 0)$ | 0.081 | 0.336 | 3.084 | 0.119 | 0.935 | 0.989 | 0.997 |
| Multi-Scale L2 $(\alpha_{un} = 1, \alpha_{dms} = 1)$ | 0.077 | 0.332 | 3.075 | 0.115 | 0.939 | 0.989 | 0.997 |
| Full DispNet-$S^4$ (perceptual [21] loss) $(\alpha_{un} = 1, \alpha_d = 0.01)$ | 0.075 | 0.316 | 3.020 | 0.112 | 0.940 | 0.990 | 0.998 |
| Baseline-CS $(\alpha_{un} = 0, \alpha_d = 0)$ | 0.136 | 0.952 | 4.434 | 0.181 | 0.848 | 0.964 | 0.986 |
| Full-CS $(\alpha_{un} = 1, \alpha_d = 0.001)$ | 0.121 | 0.662 | 4.083 | 0.161 | 0.868 | 0.974 | 0.993 |

Table 2. Ablation studies of our DispNet-$S^4$ on KITTI [7].

tween $\mathbf{D}_0$ and $\mathbf{D}_0^g$ (denoted by 'Full DispNet-$S^4$') yields the best KITTI metrics in Table 2 and the most detailed and consistent depth maps as seen in Fig. 5. Finally, to show that our $S^4$ does not only work on KITTI, we trained our DispNet-$S^4$ on the CityScapes [1] dataset, as denoted by -CS in Table 2. Again, our method with $S^4$ yields consistent improvements over the Baseline-CS model.

## 5. Discussion

While our proposed method yields quantitative results on KITTI in pair with the SOTA, it clearly generates depth maps with more precise boundaries without requiring additional inputs, as shown in Fig. 4. However, there still exists room for further improvement in terms of learning the feature extraction in our $l_d$ and using 'intensity' invariant losses [13] to only penalize structure errors without affecting the scale or shift of the depth values in our depth detail losses. Furthermore, in order to clearly observe the effects of our proposed $S^4$, we opted for a vanilla CNN with no advanced blocks, such as attention mechanisms [31], information-preserving down- and up-convolutions [15] and adaptive quantization [12] which could be further integrated to improve the SVDE performance. Also, it is worthwhile to investigate the benefits of *un*-blurring and *un*-noising in other computer vision tasks where structural details should be preserved, such as semantic segmentation and self-supervised optical flow estimation.

## 6. Conclusions

We proposed a novel self-supervised structural sharpening technique, referred to as $S^4$, to guide the SVDE networks in learning the high-level of details onto the estimated depth maps. Our $S^4$ serves as the bridge between the RGB domain and the depth domain by transferring the recovered structural details via the *un*-blurring and *un*-noising operations into the estimated depth maps by adaptive convolutions to yield structurally sharpened depths. Such sharpened depths are back again utilized for self-supervision without any ground truth depth or segmentation maps. Our network with $S^4$ qualitatively outperforms and is quantitatively at pair with previous SOTA self-supervised SVDE methods without the need for advanced or heavy network blocks and further approached the level of fully-supervised methods in terms of preserving the structural details and sharpness that are already present in the input RGB images.

## 7. Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00144444, Deep Learning Based Visual Representational Learning and Rendering of Static and Dynamic Scenes).

# References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 6, 7, 8

[2] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013. 1

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014. 2, 6, 7

[4] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5515–5524, 2016. 1

[5] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 2, 7

[6] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 6, 7, 8

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3838, 2019. 2, 3, 6, 7

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. 2, 5

[10] Juan Luis Gonzalez Bello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. In *Advances in Neural Information Processing Systems*, volume 33, pages 12626–12637. Curran Associates, Inc., 2020. 2, 3, 5

[11] Juan Luis Gonzalez Bello and Munchurl Kim. Plade-net: Towards pixel-level accuracy for self-supervised single-view depth estimation with neural positional encoding and distilled matting loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6851–6860, June 2021. 2, 3, 6, 7

[12] Juan Luis Gonzalez Bello, Jaeho Moon, and Munchurl Kim. Positional information is all you need: A novel pipeline for self-supervised svde from videos, 2022. 2, 3, 6, 7, 8

[13] Juan Luis Gonzalez Bello, Soomin Seo, and Munchurl Kim. Pan-sharpening with color-aware perceptual loss and guided re-colorization. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 908–912, 2020. 8

[14] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon. Sparse auxiliary networks for unified monocular depth prediction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11078–11088, June 2021. 2

[15] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 6, 7, 8

[16] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *International Conference on Learning Representations*, 2020. 1, 3, 7

[17] Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2019. 7

[18] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016. 1

[19] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005. 2

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4, 6, 8

[22] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12642–12652, October 2021. 3, 6, 7

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[24] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020. 3

[25] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *CoRR*, abs/1907.10326, 2019. 2, 6

[26] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 1

[27] Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala. Content-preserving warps for 3d video stabilization. In *ACM SIGGRAPH 2009 Papers*, SIGGRAPH '09, pages 44:1–44:9, New York, NY, USA, 2009. ACM. 1

[28] Qi Liu, Xinbo Gao, Lihuo He, and Wen Lu. Single image dehazing with depth-aware non-local total variation regularization. *IEEE Transactions on Image Processing*, 27(10):5178–5191, 2018. 1

[29] Yang Liu, Jinshan Pan, Jimmy Ren, and Zhixun Su. Learning deep priors for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[30] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018. 7

[31] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: high resolution self-supervised monocular depth estimation. *CoRR abs/2012.07356*, 2020. 3, 6, 7, 8

[32] S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9685–9694, June 2021. 2

[33] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017. 3

[34] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021. 2

[36] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[37] Faraz Saeedan and Stefan Roth. Boosting monocular depth with panoptic segmentation maps. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3852–3861, 2021. 2

[38] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 3

[39] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020. 2

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

[41] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019. 2

[42] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 International Conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 6, 7

[43] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7364–7373, 2019. 1

[44] Zhou 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*, 13(4):600–612, 2004. 3

[45] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2162–2171, 2019. 2, 7

[46] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1174, June 2021. 3

[47] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[48] O. J. Woodford, I. D. Reid, P. H. S. Torr, and A. W. Fitzgibbon. On new view synthesis using multiview stereo. In *In Proc. BMVC*, pages 1120–1129, 2007. 1

[49] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[50] Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. Deep view synthesis from sparse photometric images. *ACM Transactions on Graphics (ToG)*, 38(4):1–13, 2019. 1

[51] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 7

[52] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017. 2, 3

[53] Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 7