

# Topological Regularization for Dense Prediction

Deqing Fu<sup>1</sup> and Bradley J. Nelson<sup>1</sup>

<sup>1</sup>Department of Statistics, University of Chicago

{deqing, bradnelson}@uchicago.edu

**Abstract**—Dense prediction tasks such as depth perception and semantic segmentation are important applications in computer vision that have a concrete topological description in terms of partitioning an image into connected components or estimating a function with a small number of local extrema corresponding to objects in the image. We develop a form of topological regularization based on persistent homology that can be used in dense prediction tasks with these topological descriptions. Experimental results show that the output topology can also appear in the internal activations of trained neural networks which allows for a novel use of topological regularization to the internal states of neural networks during training, reducing the computational cost of the regularization. We demonstrate that this topological regularization of internal activations leads to improved convergence and test benchmarks on several problems and architectures.

**Index Terms**—Computational topology, Topological data analysis, Persistent homology, Monocular depth estimation, Semantic segmentation

## I. INTRODUCTION

Dense prediction problems are a class of problems which attempt to infer some value at every pixel in an image. Examples include semantic segmentation and monocular depth estimation which we consider here, as well as instance segmentation, hierarchical boundary detection, and figure-ground inference. Convolutional neural network models have proved to be highly effective at solving such tasks, and many common architectures for dense prediction are built using an encoder and decoder stitched together in some variation of a U-Net [1]. Despite their success, neural networks have drawbacks such as expensive training (both in terms of time and power consumption) and typically requiring large amounts of data [2]. Regularization using some form of prior knowledge about a problem can be beneficial in reducing the amount of data or the number of iterations needed to train a network, and can also improve the generalization of the network [3].

Dense prediction tasks on natural images typically ask a network to infer properties of regions of an image corresponding to individual objects. This partitions an image into relatively few components where pixels in a shared component are treated in a similar manner, at least in the output of the network. Viewing the output of the task as a function on pixels, this means we expect a function (not necessarily smooth) with few local extrema corresponding to the regions of the image that are most or least prominent in the prediction task. Total variation regularization [4] is commonly used in imaging tasks where the output is expected to be piece-wise constant. However, another natural point of view which is agnostic to

the smoothness of the function is to penalize the number of maxima or minima directly using level set topology.

We introduce a regularization method based on the super-level set topology of a function which encourages few local maxima in the output of a dense prediction problem. This method is based on persistent homology [5] and joins a growing body of topological regularization methods finding use in machine learning [6], [7], [8], [9], [10]. We find that this super-level set topology is not only important in the output of the network, but also in the internal activations of the decoder portion of the network. We demonstrate that topological regularization of the internal activations of the network during training leads to faster convergence and improved inference results on both a semantic segmentation task and a monocular depth estimation task.

## II. RELATED WORK

**Semantic Segmentation** is a common task in image processing and it requires a dense prediction on each pixel of its corresponding classification. Convolutional neural networks (CNNs) [11] have shown great capability in solving semantic segmentation problems, and many architectures and methods have been introduced. Region based methods such as R-CNN [12], and its derivative, Mask R-CNN [13] use a CNN as a feature extractor for region proposals and refine from bounding boxes to semantic segmentation masks. FCN [14] demonstrates pixel-to-pixel prediction capabilities, U-Net [1] shows great performance on medical images and DeepLab [15] proposes atrous spatial pyramid pooling (ASPP) to handle objects of multiple scales. In recent works people are also interested in finding ways of learning semantic representations in an unsupervised manner. Zhang *et al.* [16] show a way of bootstrapping semantic representation learning via primitive hierarchical grouping such as edges. Aside from development in model architectures and learning schemes, another important line of work seeks to regularize networks to produce high-quality and robust semantic results. Jia *et al.* [17] shows that integrating the total variation regularization [4] to deep convolutional neural networks can both improve the quality of segmentation and its robustness to noises. The key difference of this work to Jia *et al.*'s is that they only regularize the last softmax layer while this work uses topology to regularize internal activations, not just the last layer.

**Monocular Depth Estimation** is another topic of interest in the area of dense predictions. There is a wide range of contributions to learning depth from stereo images. The goal of monocular depth estimation, however, is to predict depth

from a single RGB image. Eigen *et al.*[18] initiated the idea of using deep neural networks to estimate depth without using superpixelation. More recent work follow the step of improving the capability of convolutional neural networks. Alhashim *et al.*[19] propose the DenseDepth model that leverage transfer learning, to use ImageNet [20] pretrained DenseNet [21] as encoders to extract features and a decoder composed of basic convolution layers to reconstruct depth. Lee *et al.*[22] proposed the BTS model that use local planar guidance layers to further improve the encoder-decoder scheme.

Aside from the supervised manner of training on one specific dataset, Lasinger *et al.*[23] demonstrate a zero-shot technique to mix multiple data sets, even with incompatible annotations, which can improve the generalization skill of the networks. Outside of the family of convolutional neural networks, Ranftl *et al.*[24] extends the vision transformers [25] to monocular depth estimation, and achieves better results than CNNs. This work will focus on CNNs, but discussions on vision transformers will be a promising direction for future work. Depth maps are intrinsically equipped with one nice property: they only have a small number of local extrema corresponding to object instances in the images. This work will leverage this property and show that a topological regularization on internal activations can also improve monocular depth estimation tasks.

**Topology and Neural Networks.** The key contribution of this work is to introduce a method of topological regularization for dense prediction problems which is built on persistent homology. Persistent homology [5] is an algebraic signature of filtrations of topological originating in topological data analysis which measures the robustness of topological features such as connected components and holes in a topological space. There has been a recent broad effort to develop persistent homology-based losses and regularization terms in a variety of contexts, including the development of application-specific losses [26], [27], [28], [29] as well as general implementations and theory to make this tool increasingly available [6], [7], [8], [9], [10]. Several applications to computer vision have previously appeared, including a previous application to semantic segmentation by Hu *et al.*[30]. In contrast to Hu *et al.* our method does not use any form of ground truth in the regularization, or an expensive Wasserstein distance computation. Our regularization scheme is similar to those suggested for the use in training generative adversarial networks on images of objects by Br  l-Gabrielsson *et al.*[7], and we leverage the implementation provided in their work.

A particularly novel aspect of our work is the application of topological regularization to the internal activations of a network, and, to the best of our knowledge, this is the first work to demonstrate the utility of doing so. Typically, topological losses or regularization terms are applied to the output of the network, but as we see, dense prediction tasks manifest similar topology in earlier activations as well. Our approach is inspired by recent work by Naitzat *et al.*[31] that has shown that trained neural networks tend to simplify topology in internal activations, although their topological construction uses Rips complexes built on entire data sets as they pass through a network, and our work focuses on level set filtrations of single inputs.

### III. TOPOLOGICAL REGULARIZATION

In this section we provide a brief introduction to the persistent homology of super-level set filtrations, which is the key construction in our topological regularization scheme. For surveys of persistent homology which provide broader context and additional detail see [32], [33], and for additional background on computing persistent homology see [34]. The topological spaces we use are all thought of as subsets of a rectangle, discretized using the Freudenthal triangulation so that the vertex set of the triangulation is arranged to correspond to the pixel grid of an image. This gives a combinatorial representation of the rectangle as a *simplicial complex*  $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2$ , consisting of a vertex set (0-simplices)  $\mathcal{X}_0$ , an edge set (1-simplices)  $\mathcal{X}_1 \subset \mathcal{X}_0 \times \mathcal{X}_0$ , and a set of triangles (2-simplices)  $\mathcal{X}_2 \subset \mathcal{X}_0 \times \mathcal{X}_0 \times \mathcal{X}_0$ . We will denote  $k$ -simplices as  $(x_0, \dots, x_k), x_i \in \mathcal{X}_0$ .

Images can either be input or output images of a neural network (with one or more channels), or the internal activations of a neural network on a particular image (typically many channels). A super-level set filtration uses a real-valued function  $f$  which takes values over each pixel, in our case we take the 2-norm of the channel values over each pixel. This function can be extended from a function on pixel values (the vertex set of  $\mathcal{X}$ ) to higher-dimensional simplices using a lower-star filtration  $f(x_0, \dots, x_k) = \min_{i=0, \dots, k} f(x_i)$ . A filtration is a nested sequence of topological spaces  $\{\mathcal{X}_a\}$ . Super-level set filtration for a function  $f$  extended to a whole simplicial complex use  $\mathcal{X}_a = f^{-1}([a, \infty))$ , which satisfies the inclusion condition  $\mathcal{X}_a \subseteq \mathcal{X}_b$  if  $a > b$ . An example image with several snapshots of the associated super-level set filtration can be seen in Fig. 1.

Persistent homology is an algebraic invariant of filtrations which summarizes how topological features such as connected components and holes appear and disappear as the filtration parameter increases. Homology is computed first starting with chain complexes  $\{C_k(\mathcal{X})\}_{k \geq 0}$  which are vector spaces with basis vectors for each  $k$ -simplex in  $\mathcal{X}$ , and connected by boundary maps  $\partial_k : C_k \rightarrow C_{k-1}$  defined (for simplicial complexes) as  $\partial_k(x_0, \dots, x_k) \mapsto \sum_i (-1)^k (x_0, \dots, \hat{x}_i, \dots, x_k)$ , where  $\hat{x}_i$  denotes the removal of the vertex  $x_i$  to obtain a  $k-1$  simplex from the  $k$ -simplex, and  $\partial_0 = 0$ . These boundary maps are differentials:  $\partial_k \partial_{k+1} = 0$  for all  $k \geq 0$ .

Homology in dimension  $k$  is the quotient vector space  $H_k = \ker \partial_k / \text{img } \partial_{k+1}$ . The rank of homology in dimension 0 counts the number of connected components of  $\mathcal{X}$ , in dimension 1 counts the number of holes, and, generally, in dimension

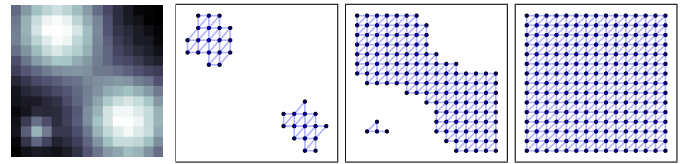


Fig. 1: Left to right: function  $f$  of pixel values over an example image where light pixels have higher value than dark. Then a sequence of simplicial complexes obtained from super-level sets:  $f^{-1}([0.85, \infty))$ ,  $f^{-1}([0.45, \infty))$ , and  $f^{-1}([0, \infty))$ .

$k$  counts  $k$ -dimensional voids. Homology is a functor, meaning that the inclusions  $\mathcal{X}_a \subseteq \mathcal{X}_b$  have associated linear maps  $F_{k;a,b} : H_k(\mathcal{X}_a) \rightarrow H_k(\mathcal{X}_b)$  which are useful in determining how features in  $\mathcal{X}_a$  map to features in  $\mathcal{X}_b$ . *Persistent homology*  $PH_k(\{\mathcal{X}_a\})$  describes how vectors first appear in the co-kernel of a map and then survive the application of maps induced by inclusion until eventually entering the kernel. Persistent homology is characterized up to isomorphism [36], [37] by birth-death pairs  $PH_k(\{\mathcal{X}_a\}) = \{(b_i, d_i)\}$  where the pair  $(b_i, d_i)$  describes the birth of a new vector at parameter  $b_i$  and death of its image at parameter  $d_i$ , often visualized plotted in the plane as a *persistence diagram* – see Fig. 2 for an example. Pairs with well-separated birth and death are robust to perturbation of function values [38], and points with nearby births and deaths are typically considered topological noise.

Each birth or death in persistent homology has a subgradient, one element of which is obtained by associating the birth or death to a particular simplex that changed the rank of homology at the parameter the birth or death occurred. While this mapping is not generally unique, algorithms for computing persistent homology produce a choice of one-to-one mapping – see Brüel-Gabrielsson *et al.* [7] for additional details. The particular regularization we use is in the family of functionals based on algebraic functions of the birth-death pairs [39]. Specifically, we penalize all but the  $k$  longest birth-death pairs in dimension 0:

$$\mathcal{L}_{\text{Topology}}(\{b_i, d_i\}) = \sum_{i > k} |d_i - b_i|^2. \quad (1)$$

This encourages at most  $k$  local maxima in the function  $f$  over the image channels.

**Computation.** To apply our topological loss we use the TopologyLayer PyTorch package [7] modified to use the union-find algorithm [40] to compute  $PH_0$ , which runs in  $O(m\alpha(m))$  time, where  $m$  is the number of edges in the simplicial complex and  $\alpha$  is the inverse Ackermann function. To compute persistent homology in higher dimensions it is necessary to perform a factorization of boundary matrices which can be achieved in matrix multiplication time [41]. Standard implementations are asymptotically cubic in the number of simplices, but sparsity in boundary matrices often

make this bound pessimistic [34]. An advantage of our method is that the size of the spaces we use to regularize internal activations of a network are an order of magnitude smaller than the size of the output space, significantly reducing the cost of using a topological penalty.

#### IV. TOPOLOGY OF INTERNAL ACTIVATIONS

In this section, we provide experimental results which demonstrate that topology can appear in the internal activations of trained neural networks. We train a convolutional neural network on binary semantic segmentation task which assigns values to each pixel indicating whether this pixel is part of a human torso or not.

**Model architecture.** We use the U-Net [1] architecture, which is a popular baseline for segmentation tasks. Our U-Net adopts 6 phases of encoders and 6 phases of decoders, where each block consists of two layers of convolutions, batch normalizations, and ReLU activations. Figure 3 shows the flow of the network, along with visualizations of each layer, and the network progresses as  $\text{Input} \rightarrow \text{enc1} \rightarrow \dots \rightarrow \text{enc6} \rightarrow \text{bottleneck} \rightarrow \text{dec6} \rightarrow \dots \rightarrow \text{dec1} \rightarrow \text{Output}$  with skip connections.

**Dataset.** We use the COCO dataset [35], specifically COCO-2014, which contains 83K training images and 41K validation images. From its provided semantic segmentation annotations, we make a subset where all images contain human annotations and further process them to only annotate pixels as human/non-human. This commonly used data set contains human objects and [42], [43] have recently studied bias in captions. However, our semantic segmentation task does not use these potentially problematic labels.

**Training objectives and hyper-parameters.** The training objective is the MSE loss

$$\mathcal{L}_{\text{MSE}}(y, \hat{y}) = \sum_{i=1}^h \sum_{j=1}^w (y_{i,j} - \hat{y}_{i,j})^2 \quad (2)$$

where  $h$  and  $w$  are the height and width of the image. We train the network for 100 epochs with initial learning rate of 0.01. We use SGD as the optimizer with learning rate decaying by half every 40 epochs.

**Experimental results.** As discussed in Section III, since each intermediate layer  $h \in \mathbb{R}^{h \times w \times c}$  has more than one channels, e.g.  $c > 1$ , there's a natural and differentiable projection  $\pi$  that send the  $c$ -dimensional vector at each pixel to a real value, in order to apply super-level set filtration and compute persistence diagrams. The projection  $\pi : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^{h \times w \times 1}$  is defined as, for each pixel  $(i, j)$ ,

$$\pi(\mathbf{h}_{i,j}) = \|\mathbf{h}_{i,j}\|_2 \quad (3)$$

We evaluate on a completely trained network by visualizing the projection, by Equation (3), of each internal activation and computing persistence diagrams. As shown in the example of Figure 3, the desired ground-truth mask is consisted of two connected components since there are two human torsos in the input image.

As we follow the steps of this trained convolutional neural network by looking at its internal activations, we can see that

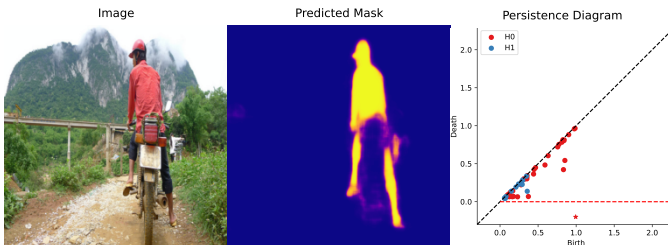
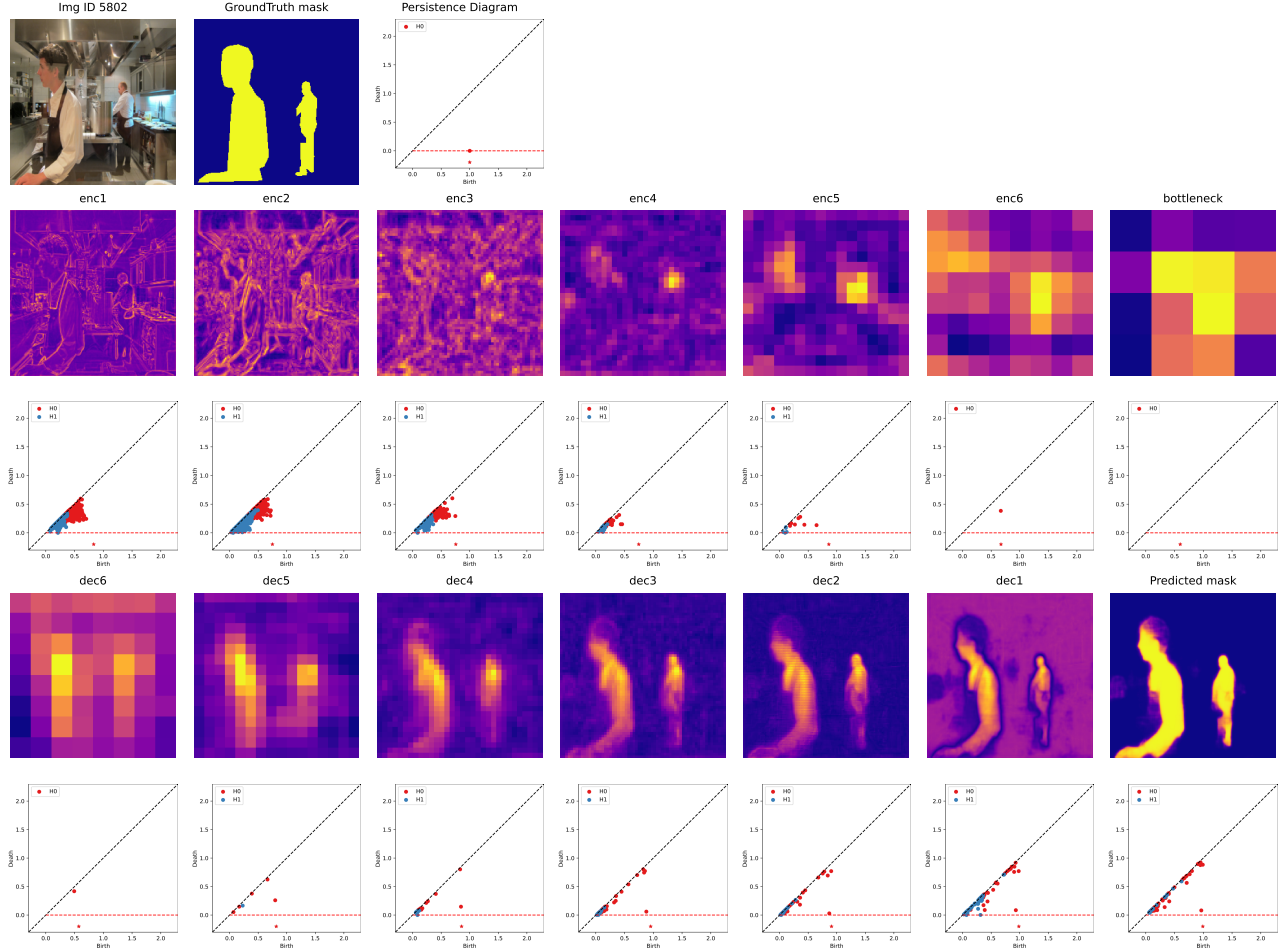


Fig. 2: Left to right: an input image from the COCO data set [35]; semantic segmentation mask; persistence diagram of the mask. Each point in the persistence diagram is a persistence pair: red points are homology in dimension 0, and blue points are homology in dimension 1. There is a single red point below the dashed red line representing the single connected component present at the end of the filtration.



**Fig. 3: Visualizations and Persistence Diagrams of Internal Activations.** The first row shows the input image, the ground truth segmentation mask, and the mask’s persistence diagram. The 2nd and 4th row show the visualizations of the magnitude of internal activations. On the 3rd and 5th row, each figure shows the persistence diagrams of the corresponding internal activations above. Points further from the diagonal are more robust features. Both visualizations and diagrams show that the level-set topology can emerge as early as in the 4th encoder layer, and it remains nearly consistent in the decoder layers. No topological regularization is used on this example.

the first 3 encoder layers, *enc1*, *enc2*, and *enc3*, present low-level edge structures, and their corresponding diagrams do not reveal any simple topology. Starting from the 4th encoder layer both the visualizations and diagrams demonstrate that topology starts to simplify and two connected components emerge seen as two  $PH_0$  pairs well separated from the diagonal.

Once we proceed to the decoder layers, both visualizations and persistence diagrams illustrate that the topology remains fairly consistent through all decoder layers and the output layer. In Figure 3 these two components eventually form the segmentation masks for the two people in the image. In other images the number of robust components in the internal activations typically agrees with the number of connected components in the final segmentation mask. More examples can be viewed in the supplemental materials. This provides a natural suggestion of explicitly regularizing the internal activations throughout the training stage.

**Topological Regularization on binary Semantic Segmentation.** We experimented using the same architecture as de-

scribed above and trained the networks for 50 epochs with cosine learning rate schedule, but varying the choice of regularization. We regularize *dec4* with  $k = 8$  to penalize more than 8 connected components on this intermediate layer. As shown in Table I, there is a improvement on mIoU, e.g., mean Intersection-over-Union accuracy, when we regularize the second decoder layer with the proposed topological regularizer. As shown in Figure 4, there are also improvements on convergence speed with the assistance of the proposed regularization.

The next section will show how this explicitly regularization can improve the convergence and test benchmarks on several architectures on depth prediction tasks.

	mIoU (%)
No Regularization	51.64
Topological Regularization	<b>52.54</b>

TABLE I: Performance Improvement by Topology Regularizer



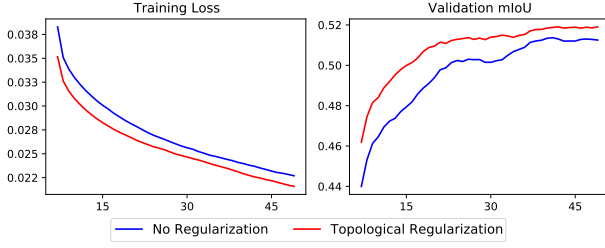


Fig. 4: Convergence Improvement by Topology Regularizer

## V. EXPERIMENTS

In this section, we explore the performance improvement on monocular depth estimation with different level of regularizations. We start with the base U-Net architecture, and compare the performance without regularization, with only total variation regularization, and with both total variation and topological regularization. We'll show that both regularization on internal activations assist the convergence and performance of the U-Net model. Subsequently, we experiment with a recent state-of-the-art architecture DenseDepth to demonstrate the versatility of our proposed regularization.

### A. Training Objectives

Our proposed training objective is a weighted sum of three loss functions and two regularization terms:

$$\begin{aligned} \mathcal{L}(y, \hat{y}) = & \lambda_d \cdot \mathcal{L}_{\text{depth}}(y, \hat{y}) + \lambda_g \cdot \mathcal{L}_{\text{gradient}}(y, \hat{y}) + \lambda_s \cdot \mathcal{L}_{\text{SSIM}}(y, \hat{y}) \\ & + \lambda_{\text{tv}} \cdot \mathcal{L}_{\text{TotalVariation}}(\hat{h}^{(a)}) + \lambda_{\text{top}} \cdot \mathcal{L}_{\text{Topology}}(\hat{h}^{(b)}) \end{aligned} \quad (4)$$

where  $y$  represents the ground-truth label of depth,  $\hat{y}$  represents the predicted depth,  $\hat{h}^{(a)}$  and  $\hat{h}^{(b)}$  represents some intermediate layers of the network. Each loss term is defined as follows, **Depth Loss.** We use the RMSE loss in log scale which empirically converges faster than L1 or L2 loss.

$$\mathcal{L}_{\text{depth}}(y, \hat{y}) = \sqrt{\sum_{i=1}^h \sum_{j=1}^w (\log y_{i,j} - \log \hat{y}_{i,j})^2} \quad (5)$$

**Gradient Loss.** The horizontal and vertical image gradients,  $\nabla_{\parallel}$  and  $\nabla_{\perp}$ , are computed by a Sobel filter [44], [45]. This further helps align the edges of ground-truths and predictions.

$$\mathcal{L}_{\text{gradient}}(y, \hat{y}) = \sum_{i=1}^h \sum_{j=1}^w (|\nabla_{\perp} y_{i,j} - \nabla_{\perp} \hat{y}_{i,j}| + |\nabla_{\parallel} y_{i,j} - \nabla_{\parallel} \hat{y}_{i,j}|) \quad (6)$$

**Structural Similarity Loss.** This loss uses the Structure Similarity Index Measure (SSIM) [46] which is shown to be a good loss term for depth estimation tasks [47].

$$\mathcal{L}_{\text{SSIM}}(y, \hat{y}) = \frac{1 - \text{SSIM}(y, \hat{y})}{2} \quad (7)$$

**Total Variation Regularization.** It's often used in imaging tasks where the expected output is piece-wise constant. We apply this regularization term to the last layer  $\hat{h}^{(a)}$  before the output layer.

$$\begin{aligned} \mathcal{L}_{\text{TotalVariation}}(\hat{h}^{(a)}) = & \sum_{k=1}^c \sum_{i=1}^{h-1} \sum_{j=1}^w (\hat{h}_{i+1,j,k}^{(a)} - \hat{h}_{i,j,k}^{(a)})^2 \\ & + \sum_{k=1}^c \sum_{i=1}^h \sum_{j=1}^{w-1} (\hat{h}_{i,j+1,k}^{(a)} - \hat{h}_{i,j,k}^{(a)})^2 \end{aligned} \quad (8)$$

**Topological Regularization.** We apply this regularizer to the second decoder layer  $\hat{h}^{(b)}$ . We first use the projection described by Equation (3) to project this internal activation to  $\tilde{h} \in \mathbb{R}^{h \times w}$  and compute its birth-death pairs  $(b_i, d_i)$  using super-level set filtration and persistent homology described in Section III. Afterwards, we formulate the loss, given these birth-death pairs, through Equation (1), e.g.,  $\mathcal{L}_{\text{Topology}}(\{b_i, d_i\}) = \sum_{i>k} (d_i - b_i)^2$ .

For our application, we choose  $k = 8$  to penalize internal activations to have more than 8 connected components or local extrema. The hyperparameter  $k$  can be chosen differently, but we choose  $k$  to obtain the best performance.

To achieve the best performance, we set the weights of these objectives as  $\lambda_d = 0.1, \lambda_g = 1.0$  and  $\lambda_s = 1.0$ . As unregularized DenseDepth to have clearer topology on internal activations than U-Net (see Figure 5), we choose different weights of  $\lambda_{\text{tv}}$  and  $\lambda_{\text{top}}$  for each. For U-Net, we set  $\lambda_{\text{tv}} = 1.0$  and  $\lambda_{\text{top}} = 0.001$ ; and for DenseDepth, we set  $\lambda_{\text{tv}} = 0.1$  and  $\lambda_{\text{top}} = 0.0001$ .

### B. Models and Training Details

**U-Net.** We use variants of a ResNet-50 backbone U-Net developed by [48] and [49]. This replaces the encoder part of the original U-Net with a ImageNet pretrained ResNet-50 [50] feature extractor. We trained the model in three settings, without regularization, with total variation regularization, and with total variation plus topological regularization. For all three training procedures, we set the initial learning rate for parameters of the decoder to 0.03 and we keep the learning rate for parameters of the pretrained ResNet encoder to be 1/10 of that of the decoder. Models are trained for 100 epochs with batch size 16, a cosine learning rate schedule, a momentum of 0.9 and a weight decay of 1e-4. For regularization settings, the total variation regularization is enforced onto the last decoder layer and the topological regularization is enforced onto the second decoder layer.

**DenseDepth.** We use DenseDepth with DenseNet-161 encoder. We trained the model in two settings, without regularization, and with total variation regularization plus topological regularization. To be comparable with scores reported in [51], we use Adam optimizer [52] with initial learning rate 0.0001,  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . Models are trained for 20 epochs with batch size 12 and a cosine learning rate schedule. For regularization settings, we keep it the same as in the U-Net experiments, e.g., the total variation regularization is enforced onto the last decoder layer and the topological regularization is enforced onto the second decoder layer.

Experiments are trained on two nVidia 2080 Ti's with 11GB memory each.

### C. Data Set and Augmentation Policy

**DIODE** is a data set that provides images, depth maps and surface normals for both indoor and outdoor scenes [51]. Data provided are captured at a resolution of  $1024 \times 768$ . It contains 8,574 indoor scenes and 16,884 outdoor scenes for training. The maximum depth captured by the sensor is 350 meters and the minimum depth is 0.6 meters. Both our models

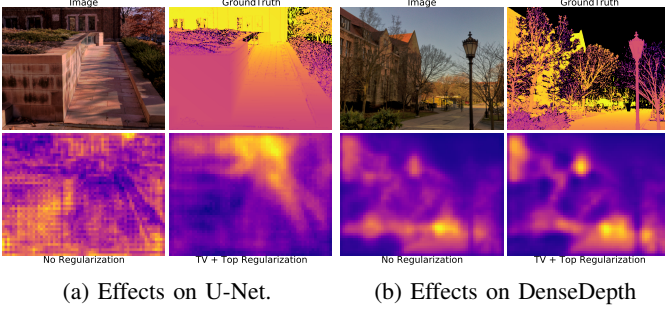


Fig. 5: **Effects of Topological Regularization on Interval Activations.** For each architecture, we visualize the second decoder layer. Our proposed regularization helps the network concentrate better on regions of interest.

take half of the original resolution as input, i.e., a resolution of  $512 \times 384$ . The U-Net model produces predictions at a resolution of  $512 \times 384$ , and the DenseDepth model produces predictions at a resolution of  $256 \times 192$  followed by a 2x upsampling through bilinear interpolations.

**Augmentation policy** Data augmentation is universally used to reduce over-fitting and can result in better generalization skills. Since the monocular depth estimation tasks aim to predict depth from an entire image, geometric transformations may not be appropriate choices since they may introduce additional distortions that are not natural in depth estimation. We adopt similar data augmentation strategies as [19]. For geometric augmentations, We only performed random horizontal flips with a probability of 0.5. For photo-metric augmentations, we performed random color jittering with a probability of 0.8, random channel swapping with a probability of 0.5, randomly converting images to gray scale with a probability of 0.2, and a random Gaussian blurring with a probability of 0.5. It’s unknown if more hand-crafted augmentations or automated augmentation policies [53], [54] can also help with generalization skills of trained networks on monocular depth estimation and is an interesting topic for future research.

#### D. Qualitative comparison

We qualitatively investigate our proposed method’s regularization effects through visualizations of the internal activations, and specifically, we visualize the second decoder layer for both trained U-Net and DenseDepth, as shown in Figure 5. In general, our proposed regularization helps the network concentrate on regions of interest at the early decoding stage.

For the U-Net without regularization, the internal activation of the trained network shown in Figure 5a concentrates evenly on nearly everywhere in the image. This might help the network gather local information but it’s not necessary for dense estimation problems. This is due to the fact that dense prediction problems tend to have region blocks whose internal gradients are minimal. Thus, it’s better for the network to concentrate on regions than on pixels during the early decoding stage, and the fine-grain information can be later accumulated through skip connections from early encoding stages. In contrast, our regularized model learns to focus on high-level regions. Our regularization also aids in reducing artifacts after unpooling and deconvolution seen in activations.

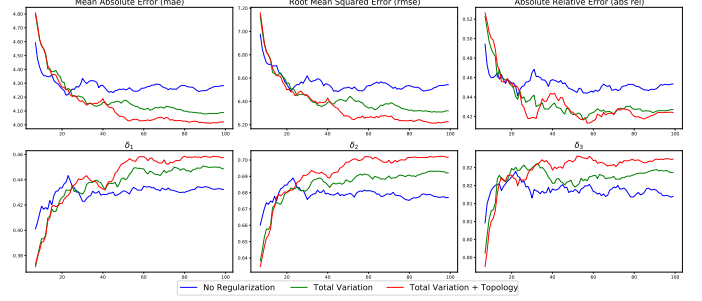


Fig. 6: **Improvements on U-Net Convergence.** Blue lines (without regularization) demonstrates some degree of overfitting as loss increases and accuracy decreases after 20 epochs. Green lines (with total variation regularization) alleviates overfitting a bit, and Red lines (with both proposed regularizations) further overcomes overfitting and converges to a better optimum.

For DenseDepth, the trained network without regularization already concentrates on regions, as shown in Figure 5b. Our proposed regularization further strengthens the level of concentration. The foreground lamps, bushes and trees tend to have lower values in our proposed method, and building regions now tend to have higher values.

In both models, we see that topological regularization aids the decoder in focusing on important regions of the image. This also provides a high level of interpretability into how successful networks are making inferences while abstracting away detail about the particular activations.

#### E. Quantitative comparison

**Benchmark Metrics.** We quantitatively compare the performance of our regularization on both U-Net and DenseDepth architectures. The accuracy metrics are defined as the following,

- i. Mean Absolute Error (mae):  $\frac{1}{hw} \sum_{i,j} |y_{i,j} - \hat{y}_{i,j}|$
- ii. Root Mean Squared Error (rmse):  $\sqrt{\frac{1}{hw} \sum_{i,j} (y_{i,j} - \hat{y}_{i,j})^2}$
- iii. Absolute Relative Error (abs rel):  $\frac{1}{hw} \sum_{i,j} \frac{|y_{i,j} - \hat{y}_{i,j}|}{\hat{y}_{i,j}}$
- iv. mae  $\log_{10}$ :  $\frac{1}{hw} \sum_{i,j} |\log_{10} y_{i,j} - \log_{10} \hat{y}_{i,j}|$
- v. rmse  $\log_{10}$ :  $\sqrt{\frac{1}{hw} \sum_{i,j} (\log_{10} y_{i,j} - \log_{10} \hat{y}_{i,j})^2}$
- vi. Threshold accuracy ( $\delta_k$ ): Percentage of pixels such that  $\max\left(\frac{y_{i,j}}{\hat{y}_{i,j}}, \frac{\hat{y}_{i,j}}{y_{i,j}}\right) = \delta_k < 1.25^k$ . We specifically care about when  $k = 1, 2$  and 3.

**Convergence Improvement.** Our proposed method is also advantageous in speed up the converge. Figure 6 plots the U-Net model’s convergence curves of both validation losses and validation threshold accuracies. The vanilla model without regularization demonstrates some degree of overfitting as the validation losses start to increase and accuracies decrease after 20 epochs. Adding a total variation regularization helps alleviate overfitting but also shows slight decrease in  $\delta_3$  accuracy after 40 epochs. Enforcing an additional topological regularization further alleviates the overfitting pattern and converges to a better optimum.

**Performance Improvement.** We benchmark our proposed method on the entire DIODE dataset of both indoor and

	Level of Regularization	lower is better					higher is better		
		mae	rmse	abs rel	mae log <sub>10</sub>	rmse log <sub>10</sub>	$\delta_1$	$\delta_2$	$\delta_3$
U-Net	None	4.2776	6.5386	0.4524	0.1706	0.2181	0.4336	0.6778	0.8128
	Total Variation	4.0952	6.3145	0.4311	0.1649	0.2103	0.4455	0.6915	0.8184
	Topology	4.0548	6.2168	0.4206	0.1651	0.2121	0.4388	0.6914	0.8295
	TV + Topology	4.0138	6.2044	0.4269	0.1614	0.2069	0.4565	0.7020	0.8232
DenseDepth	None	3.6554	5.9900	0.3648	0.1660	0.2452	0.5088	0.7481	0.8625
	Total Variation	3.5073	5.5763	0.3922	0.1427	0.1884	0.5151	0.7444	0.8665
	Topology	3.5857	5.7030	0.3921	0.1435	0.1887	0.5006	0.7418	0.8668
	TV + Topology	3.4065	5.4196	0.3908	0.1395	0.1849	0.5197	0.7582	0.8745

TABLE II: **Quantitative Comparison.** UNet and DenseDepth Performance on DIODE. Each model is compared internally with varying regularization choices. Numbers in Red indicate the best score whereas in Blue shows the second best.

outdoor scenes. The benchmark results are listed for both models, U-Net and DenseDepth in Table II. Generally, our regularized model can achieve better performance on nearly every metric, compared with the unregularized model.

For U-Net, we further study the effects of different levels of regularization. By simply adding a total variation regularization, we can already reduce the losses and increase the threshold accuracy. As we subsequently add the proposed topological regularization, the scores improve further. In terms of threshold accuracy, we can improve about 2.3% in  $\delta_1$ , 2.5% in  $\delta_2$  and 1.0% in  $\delta_3$ .

We further develop the same regularization a recent state-of-the-art network, DenseDepth, and our proposed method can also help improve benchmark scores. In terms of threshold accuracy, we can improve about 1.1% in  $\delta_1$ , 1.0% in  $\delta_2$  and 1.2% in  $\delta_3$ . For validation losses, the only metric our proposed method performs worse on is the absolute relative error. This may due to the fact that our regularization is based on super-level set filtration, which will naturally focus on high-value regions, and in depth estimation problems, these regions would be those at the background. In this sense, our proposed method may perform slightly worse on foreground objects than the baseline model without regularization.

## VI. CONCLUSION

We have shown how super-level set topology plays an important role in semantic segmentation and monocular depth estimation problems. This offers a high level of interpretability into the internal working of neural networks trained to solve these problems, and we show this can be used to regularize training for faster convergence and increased accuracy. We anticipate these insights will be applicable to other dense prediction problems, and our topological regularization techniques may be adapted to a variety of architectures.

Avenues for future work may include extending topological regularizations to videos where consecutive frames are likely to share similar topological structures. Although this work demonstrates the existence of topological structures in learned convolutional neural networks and shows a way of regularizing internal activations, whether the phenomenon and regularization are also valid for the paradigm of vision transformers (ViT) [25] is unknown and of future interest.

We believe that describing the behavior of internal activations of neural networks using topology has great potential for

explaining how neural networks operate in practice. Persistent homology has the advantage of abstracting away details about the individual weights and activations in the network while retaining important geometric information – in our case the prominence of local maxima. Regularization is one method of leveraging insights from topology to improve neural networks which we have applied to dense prediction. There are also efforts to use topology to initialize weights [55] and inform decisions in network architecture [31] which may also be applicable to dense predictions.

## ACKNOWLEDGEMENTS

DF would like to thank Michael Maire for discussions and feedback. BN was supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112190040. We thank the Department of Computer Science, University of Chicago and Toyota Technological Institute at Chicago for providing cluster resources.

## REFERENCES

- [1] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Springer International Publishing, 2015, pp. 234–241.
- [2] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, “Green AI,” *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, Nov. 2020. [Online]. Available: <http://doi.org/10.1145/3381831>
- [3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [4] E. J. Candès, J. K. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” *IEEE Transactions on Information Theory*, vol. 52, pp. 489–509, 2006.
- [5] H. Edelsbrunner, D. Letscher, and A. Zomorodian, “Topological persistence and simplification,” in *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE, 2000, pp. 454–463.
- [6] J. Leygonie, S. Oudot, and U. Tillmann, “A framework for differential calculus on persistence barcodes,” *Foundations of Computational Mathematics*, pp. 1–63, 07 2021.
- [7] R. Brüel-Gabrielsson, B. J. Nelson, A. Dwaraknath, P. Skraba, L. J. Guibas, and G. Carlsson, “A topology layer for machine learning,” in *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020. [Online]. Available: <http://arxiv.org/abs/1905.12200>
- [8] M. Carrière, F. Chazal, Y. Ike, T. Lacombe, M. Royer, and Y. Umeda, “PersLay: A neural network layer for persistence diagrams and new graph topological signatures,” in *AISTATS*, 2020. [Online]. Available: <http://arxiv.org/abs/1904.09378>
- [9] K. Kim, J. Kim, M. Zaheer, J. S. Kim, F. Chazal, and L. Wasserman, “PLLay: Efficient topological layer based on persistence landscapes,” in *Neural Information Processing Systems (NeurIPS)*, 2020.



- [10] C. D. Hofer, R. Kwitt, and M. Niethammer, "Learning Representations of Persistence Barcodes," *Journal of Machine Learning Research*, vol. 20, no. 126, pp. 1–45, 2019. [Online]. Available: <http://jmlr.org/papers/v20/18-358.html>
- [11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," 1998.
- [12] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.
- [13] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 386–397, 2020.
- [14] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640–651, 2017.
- [15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2018.
- [16] X. Zhang and M. Maire, "Self-supervised visual representation learning from hierarchical grouping," in *Neural Information Processing Systems (NeurIPS)*, 2020.
- [17] F. Jia, J. Liu, and X. Tai, "A regularized convolutional neural network for semantic image segmentation," *ArXiv*, vol. abs/1907.05287, 2019.
- [18] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014.
- [19] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *ArXiv*, vol. abs/1812.11941, 2018.
- [20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.
- [21] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.
- [22] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *ArXiv*, vol. abs/1907.10326, 2019.
- [23] K. Lasinger, R. Ranftl, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2020.
- [24] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *Int. Conf. Comput. Vis. (ICCV)*, 2021.
- [25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2021.
- [26] R. Brüel-Gabrielsson, V. Ganapathi-Subramanian, P. Skraba, and L. Guibas, "Topology-aware surface reconstruction for point clouds," *Computer Graphics Forum*, vol. 39, 2020.
- [27] C. Chen, X. Ni, Q. Bai, and Y. Wang, "A topological regularizer for classifiers via persistent homology," in *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. [Online]. Available: <http://arxiv.org/abs/1806.10714>
- [28] C. Hofer, F. Graf, B. Rieck, M. Niethammer, and R. Kwitt, "Graph Filtration Learning," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 4314–4323, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v119/hofer20b.html>
- [29] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. W. Pluim, U. Bauer, and B. H. Menze, "cDice - a Novel Topology-Preserving Loss Function for Tubular Structure Segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 16555–16564. [Online]. Available: <https://ieeexplore.ieee.org/document/9578225/>
- [30] X. Hu, F. Li, D. Samaras, and C. Chen, "Topology-Preserving Deep Image Segmentation," in *Neural Information Processing Systems (NeurIPS)*, 2019, p. 12.
- [31] G. Naitzat, A. Zhitnikov, and L.-H. Lim, "Topology of deep neural networks," *J. Mach. Learn. Res.*, vol. 21, no. 345, pp. 1–40.
- [32] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009. [Online]. Available: <http://www.ams.org/journal-getitem?pii=S0273-0979-09-01249-X>
- [33] —, "Topological pattern recognition for point cloud data," *Acta Numerica*, vol. 23, pp. 289–368, May 2014. [Online]. Available: [http://www.journals.cambridge.org/abstract\\_S0962492914000051](http://www.journals.cambridge.org/abstract_S0962492914000051)
- [34] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, "A roadmap for the computation of persistent homology," *EPJ Data Science*, vol. 6, no. 1, 2017. [Online]. Available: <http://arxiv.org/abs/1506.08903>
- [35] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.
- [36] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete & Computational Geometry*, vol. 33, no. 2, pp. 249–274, 2005. [Online]. Available: <https://doi.org/10.1007/s00454-004-1146-y>
- [37] G. Carlsson and V. de Silva, "Zigzag persistence," *Foundations of Computational Mathematics*, vol. 10, no. 4, pp. 367–405, 2010. [Online]. Available: <http://link.springer.com/10.1007/s10208-010-9066-0>
- [38] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, "Stability of Persistence Diagrams," *Discrete & Computational Geometry*, vol. 37, no. 1, pp. 103–120, Jan. 2007.
- [39] A. Adcock, E. Carlsson, and G. Carlsson, "The ring of algebraic functions on persistence bar codes," *Homology, Homotopy and Applications*, vol. 18, no. 1, pp. 381–402, 2016. [Online]. Available: <http://www.intlpress.com/site/pub/pages/journals/items/hha/content/vols/0018/0001/a021/>
- [40] R. E. Tarjan, "A class of algorithms which require nonlinear time to maintain disjoint sets," *Journal of Computer and System Sciences*, vol. 18, no. 2, pp. 110–127, Apr. 1979. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/002200079900424>
- [41] N. Milosavljević, D. Morozov, and P. Skraba, "Zigzag persistent homology in matrix multiplication time," in *Proceedings of the 27th annual ACM symposium on Computational geometry - SoCG '11*. ACM Press, 2011, p. 216. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=1998196.1998229>
- [42] D. Zhao, A. Wang, and O. Russakovsky, "Understanding and evaluating racial biases in image captioning," in *International Conference on Computer Vision (ICCV)*, 2021.
- [43] Z. Hu and J. Strout, "Exploring stereotypes and biased data with the crowd," *ArXiv*, vol. abs/1801.03261, 2018.
- [44] I. Sobel and G. M. Feldman, "An isotropic 3x3 image gradient operator," 1990.
- [45] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the sobel operator," *IEEE Journal of solid-state circuits*, vol. 23, no. 2, pp. 358–367, 1988.
- [46] 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, pp. 600–612, 2004.
- [47] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6602–6611, 2017.
- [48] K. Mate, "Backboned unet," <https://github.com/mkisantal/backboned-unet>, 2018.
- [49] P. Yakubovskiy, "Segmentation models," [https://github.com/qubvel/segmentation\\_models](https://github.com/qubvel/segmentation_models), 2019.
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [51] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, and G. Shakhnarovich, "DIODE: A Dense Indoor and Outdoor DEpth Dataset," *CoRR*, vol. abs/1908.00463, 2019. [Online]. Available: <http://arxiv.org/abs/1908.00463>
- [52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.
- [53] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3008–3017, 2020.
- [54] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, 2019.
- [55] R. Brüel-Gabrielsson and G. Carlsson, "Exposition and interpretation of the topology of neural networks," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 1069–1076.