Global-Context Augmentation and Geometric Regularization for Monocular Depth Estimation

Konstantinos Fotopoulos [*1] **Alexander Fluck** [*1] **Omar Abdesslem** [*1] **Paul Marchi** [*1]

Abstract

Monocular depth estimation is a dense prediction task requiring both local precision and global contextual reasoning, while remaining consistent with the underlying scene geometry. In this work, we investigate two complementary directions for improving monocular depth estimation: i) lightweight global-context augmentation and ii) geometric self-consistency regularization. For (i), we augment pretrained CNN backbones with parallel global-context branches incorporating dilated convolutions and spatial attention, enabling wider contextual reasoning without the computational overhead of Vision Transformers while preserving the benefits of transfer learning. For (ii), we introduce a geometric regularization objective based on differentiable view synthesis, encouraging predicted depth maps to remain self-consistent under camera transformations. Through systematic ablations, we show that augmentation slightly improves performance, while geometric self-consistency regularization provides larger and complementary gains, particularly when combined with knowledge distillation.

1. Introduction and Related Work

Depth estimation is a fundamental computer vision task with applications in robotics, navigation, autonomous driving, augmented reality, object detection, and semantic segmentation (Hu et al., 2012; Du et al., 2020; Park et al., 2017). While stereo and multi-view methods can recover depth through geometric constraints and feature correspondences (Zou & Li, 2010; Ullman, 1979), monocular depth estimation (MDE) aims to predict dense depth maps from a single image, making it significantly more ill-posed and reliant on semantic scene understanding (Eigen et al., 2014).

[*]Equal contribution [1]Department of Computer Science, ETH Zurich, Switzerland. Correspondence to: KF <kfotopoulos@student.ethz.ch>, AF <afluck@student.ethz.ch>, OA <oabdesslem@student.ethz.ch>, PM <pmarchi@student.ethz.ch>.

Computational Intelligence Lab FS2026, ETH Zuerich

Recent deep learning approaches tackle MDE as a dense prediction task using encoder-decoder or U-Net CNN architectures (Ronneberger et al., 2015). Since accurate depth prediction requires both local precision and globally consistent scene understanding, capturing long-range dependencies is particularly important. Several works have therefore explored wider-context mechanisms used in dense prediction tasks, such as dilated convolutions (Chen et al., 2017).

More recently, Vision Transformers (ViTs) have shown strong performance in MDE by modeling global relations via attention mechanisms (Ranftl et al., 2021; Zhao et al., 2022). However, ViT-based methods are computationally expensive and often require large-scale pretraining. By contrast, convolutional attention modules such as CBAM (Woo et al., 2018) offer a more efficient alternative. While such modules have been successfully applied to other dense prediction and generation tasks (Guo et al., 2022; Zhang et al., 2019), their role in MDE remains less explored, despite the importance of long-range contextual reasoning.

Another important research direction concerns training supervision. Supervised methods rely on dense ground-truth depth annotations obtained from LiDAR or RGB-D sensors (Eigen et al., 2014; Laina et al., 2016), but standard regression losses do not explicitly enforce geometric consistency. Self-supervised approaches (Godard et al., 2017; 2019) instead leverage stereo image pairs or monocular video sequences together with differentiable view synthesis (Flynn et al., 2019). In these methods, the predicted depth map is used to reconstruct a target image under known camera motion, without requiring dense depth labels. However, such approaches typically assume access to multiple views during training.

Contributions. We investigate two complementary directions for improving MDE. First, we study lightweight global-context augmentation of pretrained CNN backbones via dilated convolutions and convolutional attention, avoiding the computational cost of ViTs while benefiting from transfer-learning. Second, we propose a geometric regularization objective based on differentiable view synthesis, encouraging self-consistency of predicted depth maps under camera transformations even in supervised monocular settings lacking multiple views. Ablations show that ar-

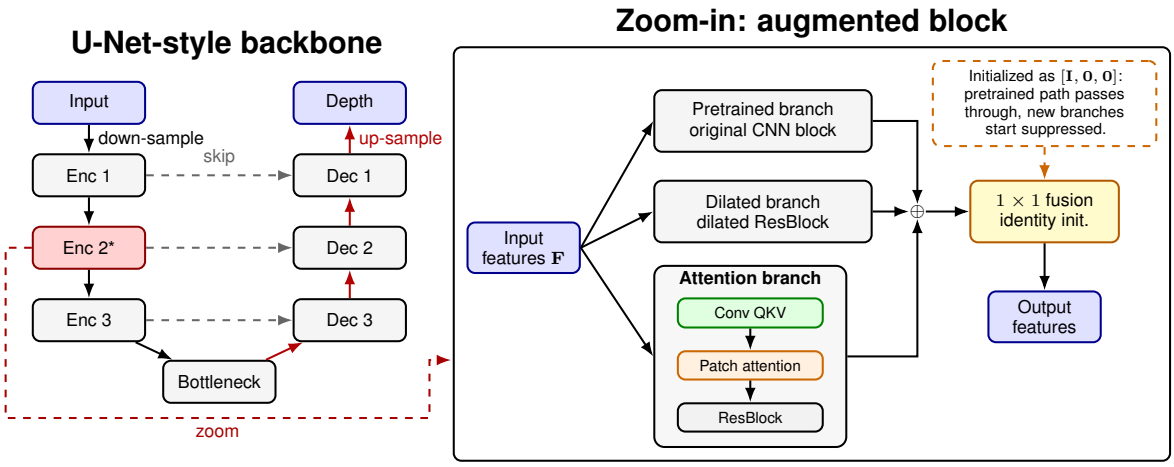


Figure 1. Proposed U-Net augmentation strategy. Selected encoder stages are expanded with parallel dilated-convolution and convolutional-attention branches. Their outputs are fused through an identity-initialized convolution, allowing the pretrained backbone to be preserved at initialization while progressively incorporating global-context features.

chitectural augmentation aids and geometric regularization improves performance, the latter providing complementary gains when combined with knowledge distillation.

2. Models and Methods

Notation. We denote by $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ an input RGB image and by $\mathbf{D} \in \mathbb{R}^{H\times W}$ its predicted depth map. Convolutional kernels are represented by $\mathbf{W} \in \mathbb{R}^{C_{\text{out}}\times C_{\text{in}}\times k\times k}$. Camera intrinsics are denoted by $\mathbf{K}$, and relative camera motion by a rigid transformation $\mathbf{T} \in \mathrm{SE}(3)$. $\|\mathbf{x}\|_\epsilon := \sum_i \sqrt{x_i^2 + \epsilon^2}$ is the Charbonnier loss.

Baseline architecture. As the baseline model, we employ a U-Net-style encoder-decoder architecture commonly used in dense prediction tasks. The encoder extracts hierarchical feature representations while reducing spatial resolution (either with max-pooling or with strided convolutions), whereas the decoder gradually restores resolution through upsampling with transposed convolutions and skip connections that preserve spatial detail. To leverage transfer learning, the encoder is initialized from pretrained CNN backbones, and decoder stages are added to construct the full U-Net architecture for dense MDE.

Global-context augmentation. To improve long-range contextual reasoning while preserving the efficiency and transfer-learning of CNN backbones, we augment selected encoder stages with parallel global-context branches, as depicted in Fig. 1. Each augmented block combines the original pretrained CNN pathway with lightweight global-context branches consisting of dilated convolutions and spatial attention. Dilated branches model medium-range spatial interactions by enlarging the receptive field while preserving a convolutional inductive bias. In contrast, the attention branch allows each spatial patch to dynamically select the regions from which information should be aggregated.

Unlike ViTs, which maintain token embeddings throughout the network, our attention branch must exchange information with convolutional branches at every layer. We therefore implement attention directly on feature maps using convolutional projections $\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V}$, producing query, key, and value feature maps $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$. These are unfolded patch-wise, and attention weights are computed from normalized query-key inner products followed by softmax normalization. The resulting weights aggregate value patches while preserving a feature-map representation. The attended features are then refined through a convolutional residual block rather than a transformer MLP.

The outputs of the pretrained, dilated, and attention branches are concatenated along the channel dimension and fused through a 1×1 convolution $\mathbf{W}_{\text{fuse}} \in \mathbb{R}^{C_{\text{out}}\times(BC_{\text{out}})\times 1\times 1}$, where B denotes the number of branches. To preserve the pretrained model at initialization, the fusion layer is initialized such that only the pretrained branch is initially propagated; the newly introduced branches are progressively incorporated during training. Formally, we initialize:

$$W_{\text{fuse}}[i, j, 0, 0] = \begin{cases} 1, & j = i \text{ and } j \in \mathcal{P}, \\ 0, & \text{otherwise}, \end{cases}$$

where $\mathcal{P}$ denotes the pretrained branch's channel indices.

Geometric regularization. We introduce a geometric self-consistency regularization objective to encourage synthesized views generated through different transformation paths to remain mutually consistent. Starting from an input image $\mathbf{I}_0$, we compare a target view synthesized directly through a transformation $0 \to 1$ against the same target view synthesized through a composed path $0 \to a \to 1$. Since MDE is inherently ambiguous, this objective alone does not uniquely

determine correct depth maps: for instance, a constant depth map remains self-consistent under view synthesis. We therefore use the proposed objective only as a regularizer on top of supervised depth losses, acting as a geometric prior encouraging stable predictions under viewpoint changes.

Let $\pi(\mathbf{x}) = \mathbf{K}\mathbf{x}/x_z$ denote projection from 3D camera coordinates to pixels, and let $\pi^{-1}(\mathbf{u}, d) = d\,\mathbf{K}^{-1}\bar{\mathbf{u}}$ denote back-projection of homogeneous pixel $\bar{\mathbf{u}} = (u, v, 1)^\top$ with depth d. Given a source image $\mathbf{I}_s$, a depth map $\mathbf{D}_t$ in the target configuration, and a transform $\mathbf{T}_{t\to s}$, inverse view synthesis samples the source image at

$$\mathbf{u}_s = \pi\left(\mathbf{T}_{t\to s}\pi^{-1}(\mathbf{u}_t, \mathbf{D}_t(\mathbf{u}_t))\right), \quad \hat{\mathbf{I}}_t(\mathbf{u}_t) = \mathbf{I}_s(\mathbf{u}_s),$$

where $\mathbf{I}_s(\mathbf{u}_s)$ is obtained through bilinear interpolation. The inverse warp additionally produces a validity mask $\mathbf{M}$ selecting pixels whose projected source coordinates lie inside the image:

$$(\hat{\mathbf{I}}_t, \mathbf{M}) = \mathcal{S}(\mathbf{I}_s, \mathbf{D}_t, \mathbf{T}_{t\to s}).$$

Given an input image $\mathbf{I}_0$, we first predict a depth map

$$\mathbf{D}_0 = f(\mathbf{I}_0; \theta),$$

which is detached for stability when synthesizing an intermediate view:

$$(\mathbf{I}_a, \mathbf{M}_a) = \mathcal{S}(\mathbf{I}_0, \mathrm{sg}(\mathbf{D}_0), \mathbf{T}_{a\to 0}).$$

We then infer $\mathbf{D}_a = f(\mathbf{I}_a; \theta)$ and synthesize the target image through the composed path:

$$(\mathbf{I}_1^{(a)}, \mathbf{M}_1^{(a)}) = \mathcal{S}(\mathbf{I}_a, \mathbf{D}_a, \mathbf{T}_{1\to a}).$$

In parallel, we synthesize the same target view directly from the source:

$$(\mathbf{I}_1^{(0)}, \mathbf{M}_1^{(0)}) = \mathcal{S}(\mathbf{I}_0, \mathrm{sg}(\mathbf{D}_0), \mathbf{T}_{1\to 0}).$$

The geometric regularization loss compares the two synthesized target views over their common valid region $\mathbf{M} = \mathbf{M}_a^{(0)} \cap \mathbf{M}_1^{(a)} \cap \mathbf{M}_1^{(0)}$:

$$\mathcal{L}_{\mathrm{sc}} = \|\mathbf{M}\odot(\mathbf{I}_1^{(a)} - \mathbf{I}_1^{(0)})\|_\epsilon + \lambda_{\mathrm{edge}}\|\mathbf{M}\odot(\nabla\mathbf{I}_1^{(a)} - \nabla\mathbf{I}_1^{(0)})\|_1.$$

In our experiments, $\lambda_{\mathrm{edge}} = 0.5$ and $\mathbf{T}_{a\to 0}$ and $\mathbf{T}_{1\to a}$ are sampled as random translations along the x- and z-axes.

3. Experimental Results

Experimental setup. We train on the provided dataset of $22,605$ RGB-depth pairs, bilinearly resized to 384×384 to facilitate more runs, using an $85/15$ random train/validation split. Ground-truth depth is clipped to $[0^+, 80]$m and normalized to $[0^+, 1]$. The training objective combines a supervised ground-truth (GT) loss with optional knowledge distillation (KD) and self-consistency (SC) regularizers:

$$\mathcal{L} = \mathcal{L}_{\mathrm{gt}} + \lambda_{\mathrm{kd}}\mathcal{L}_{\mathrm{kd}} + \lambda_{\mathrm{sc}}\mathcal{L}_{\mathrm{sc}}.$$

We define the SILog-MSE loss

$$\delta_{ij} = \log(D_{\mathrm{pred}}[i,j]) - \log(D_{\mathrm{r}}[i,j]),$$

$$\mathcal{L}_{\mathrm{sil}}(\mathbf{D}_{\mathrm{pred}}, \mathbf{D}_{\mathrm{r}}; \lambda) = \frac{1}{HW}\sum_{i,j}\delta_{ij}^2 - \lambda\left(\frac{1}{HW}\sum_{i,j}\delta_{ij}\right)^2$$

and GT and KD losses are then given by

$$\mathcal{L}_{\mathrm{gt}}(\mathbf{D}_{\mathrm{pred}}) = \mathcal{L}_{\mathrm{sil}}(\mathbf{D}_{\mathrm{pred}}, \mathbf{D}_{\mathrm{gt}}; 0.5),$$
$$\mathcal{L}_{\mathrm{kd}}(\mathbf{D}_{\mathrm{pred}}) = \mathcal{L}_{\mathrm{sil}}(\mathbf{D}_{\mathrm{pred}}, \mathbf{D}_{\mathrm{teacher}}; 0.5).$$

Validation uses SILog-RMSE $\sqrt{\mathcal{L}_{\mathrm{sil}}(\mathbf{D}_{\mathrm{pred}}, \mathbf{D}_{\mathrm{gt}}; 1)}$ to match the grading objective. All configurations are trained for 25 epochs with AdamW (Loshchilov & Hutter, 2017) (lr=10^{-4}, weight decay=10^{-4}), batch size 8, mixed-precision, and gradient-norm clipping at 1.0. KD runs use MiDaS-small (Ranftl et al., 2020) as teacher with λ_{kd}=0.2, with predictions precomputed once and per-image min/max-rescaled to align with ground-truth scale. SC runs use λ_{sc}=0.1, with $\mathbf{T}_{a\to 0}$ and $\mathbf{T}_{1\to a}$ randomly sampled as translations with $|t_x|, |t_z| \in [0.001, 0.005]$. Pretrained branches use ImageNet-pretrained, PyTorch-available VGG-16 and ResNet-18 (Simonyan & Zisserman, 2014; He et al., 2016) layers. All configurations were evaluated with the same seeds. To establish statistical significance, results report mean and std across 5 seeds.

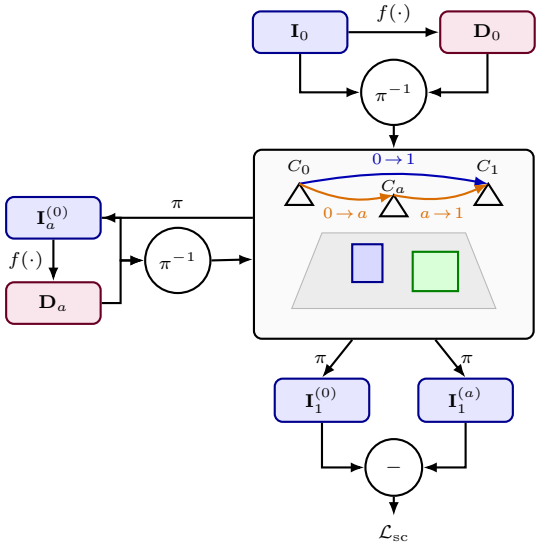


Figure 2. Geometric self-consistency regularization. A predicted depth map is used to synthesize a target view through a direct path $0 \to 1$ and a composed path $0 \to a \to 1$. The two target images are compared with a photometric and edge-aware loss. Camera displacements are exaggerated for clarity.

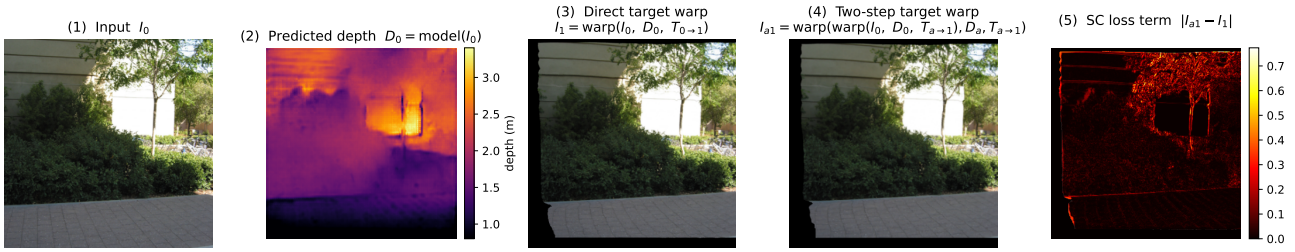


Figure 3. Visualization of the SC loss pipeline. (1): Input image, (2): predicted depth, (3, 4): direct and two-step views resp., (5): SC loss.

Table 1. Test (Kaggle) SILog-RMSE and relative differences (lower are better) for all configurations. VGG=VGG-16, RN18 = ResNet-18, D = dilation, A = attention.

Configuration	Test loss ↓	Diff. (rel.) ↓
VGG	$0.686_{\pm 0.017}$	$+0.114_{\pm 0.019}$
RN18	$0.572_{\pm 0.011}$	0.0
+ D	$0.562_{\pm 0.006}$	$-0.010_{\pm 0.010}$
+ D + A	$0.564_{\pm 0.004}$	$-0.007_{\pm 0.007}$
+ D + A + KD	$0.547_{\pm 0.002}$	$-0.025_{\pm 0.009}$
+ D + A + SC	$0.546_{\pm 0.004}$	$-0.025_{\pm 0.013}$
+ D + A + KD + SC	$\mathbf{0.537_{\pm 0.004}}$	$\mathbf{-0.034_{\pm 0.011}}$

Augmentation. We ablate the effect of architectural augmentation. Results are reported in Table 1 (first half). Consistent with Laina et al. (2016), the ResNet-18-based U-Net substantially outperforms the VGG-16-based one. Augmenting the ResNet-18 model with dilated convolutions, with or without attention, yields a modest reduction in test loss. However, these gains are small relative to the observed variability across seeds, and the attention-augmented variant does not outperform the dilation-only model.

Geometric regularization. We further ablate the effect of geometric self-consistency (SC) regularization. For these experiments, we use the ResNet-18 U-Net augmented with dilation and attention and train it with different combinations of ground-truth supervision, knowledge distillation (KD), and SC regularization. Results are summarized in Table 1 (second half). Both KD and SC reduce the test loss relative to the non-regularized architecture, with SC achieving gains comparable to KD despite not relying on a teacher. Combining KD and SC yields the best overall performance, reducing the test loss from $0.564_{\pm 0.004}$ to $0.537_{\pm 0.004}$.

Qualitative results. In Fig. 3 we also present a qualitative example of the SC loss pipeline. Predicted depth maps are used to construct a direct and two-step view synthesis. The SC loss catches inconsistencies in the two synthetic views.

4. Discussion

Architectural augmentation yielded only modest gains. While both dilation-based variants reduced the mean test error relative to the plain ResNet-18 U-Net, the improvements were small. Adding attention on top of dilation provided no further benefit. One possible explanation is that depth estimation is governed by relative geometric relationships: once information can propagate between neighboring objects, long-range dependencies may be captured transitively. Under this view, the enlarged receptive field provided by dilation may be sufficient, limiting the utility of additional attention. Alternatively, the convolutional attention mechanism may have been too weak or insufficiently optimized to exploit global context, as attention-based modules often require smaller learning-rate schedules than convolutional layers.

The proposed self-consistency (SC) regularizer relies on simplifying assumptions. To avoid the cost of forward warping, we employed inverse warping and approximated target-view depths using source-view predictions, assuming small error for the chosen displacements. Furthermore, occlusions, disocclusions, reflections, and scale inconsistencies were not explicitly modelled. Despite these limitations, SC achieved gains comparable to knowledge distillation (KD) while requiring no teacher model. Including all proposed components yields the best result, reducing the test SILog-RMSE from $0.572_{\pm 0.011}$ to $0.537_{\pm 0.004}$. Although the effect of any single component is limited, all consistently shift the mean in the same direction, resulting in a cumulative improvement that exceeds the observed randomness. Use of more elaborate view synthesis (e.g., as in (Godard et al., 2017; 2019)), is an interesting direction for future work.

5. Conclusion and Summary

We investigated architectural augmentation and geometric self-consistency (SC) regularization for monocular depth estimation. While augmentation provided only modest gains, SC achieved improvements comparable to knowledge distillation despite requiring no teacher model. Combining all proposed components yielded the best performance, reducing the test SILog-RMSE from (0.572) to (0.537). These results suggest that SC provides a useful and complementary supervision signal. Overall, this work highlights that geometry-informed training can improve depth estimation even in purely monocular training settings.

Broader Impact Statement

Monocular depth estimation can benefit applications such as robotics and autonomous navigation. However, inaccurate depth predictions may lead to unsafe downstream decisions. Such systems should be thoroughly validated and not relied upon as the sole source of information in safety-critical applications.

Declaration of AI Usage

AI tools were used as implementation and writing assistants throughout the project.

The project conception, literature review, model design, experimental design, ablation studies, loss formulation, evaluation protocol, and interpretation of results were developed by the authors. AI tools were not used to generate research ideas, select the final approach, or autonomously conduct experiments.

For implementation, AI was primarily used to translate clearly specified functionality into Python code, including model definitions, utility functions, training and evaluation pipelines, debugging assistance, and boilerplate code. The authors provided the architectural and functional specifications for the requested code and reviewed, modified, and validated the generated implementations.

AI was also used during repository refactoring to transform an initially notebook-based implementation into a modular codebase suitable for running larger-scale experiments.

For report preparation, AI was used for language editing, phrasing suggestions, and improving clarity of exposition. The technical content, experimental results, interpretations, and conclusions were written by the authors.

Acknowledgments

KF acknowledges financial support for his master's studies through a scholarship from the Union of Greek Shipowners.

References

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Du, R., Turner, E., Dzitsiuk, M., Prasso, L., Duarte, I., Dourgarian, J., Afonso, J., Pascoal, J., Gladstone, J., Cruces, N., et al. Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 829–843, 2020.

Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Flynn, J., Broxton, M., Debevec, P., DuVall, M., Fyffe, G., Overbeck, R., Snavely, N., and Tucker, R. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2367–2376, 2019.

Godard, C., Mac Aodha, O., and Brostow, G. J. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 270–279, 2017.

Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., and Hu, S.-M. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in neural information processing systems*, 35:1140–1156, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hu, G., Huang, S., Zhao, L., Alempijevic, A., and Dissanayake, G. A robust rgb-d slam algorithm. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1714–1719. IEEE, 2012.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pp. 239–248. IEEE, 2016.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Park, S.-J., Hong, K.-S., and Lee, S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 4980–4989, 2017.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

Ranftl, R., Bochkovskiy, A., and Koltun, V. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ullman, S. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019.

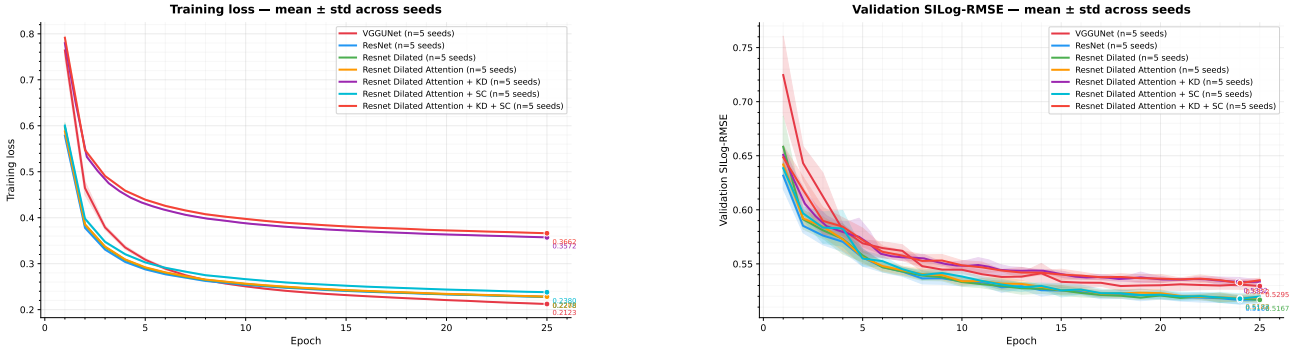
Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., and Mattoccia, S. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 international conference on 3D vision (3DV)*, pp. 668–678. IEEE, 2022.

Zou, L. and Li, Y. A method of stereo vision matching based on opencv. In *2010 International conference on audio, language and image processing*, pp. 185–190. IEEE, 2010.

A. Supplementary Experimental Results

We provide additional results regarding: (i) the training stability of all proposed configurations, and (ii) relative improvements between all pairs of configurations. For (i), note that configurations trained with different objectives have different loss definitions. Consequently, their training loss curves should be interpreted only within a given configuration and are not directly comparable across configurations.

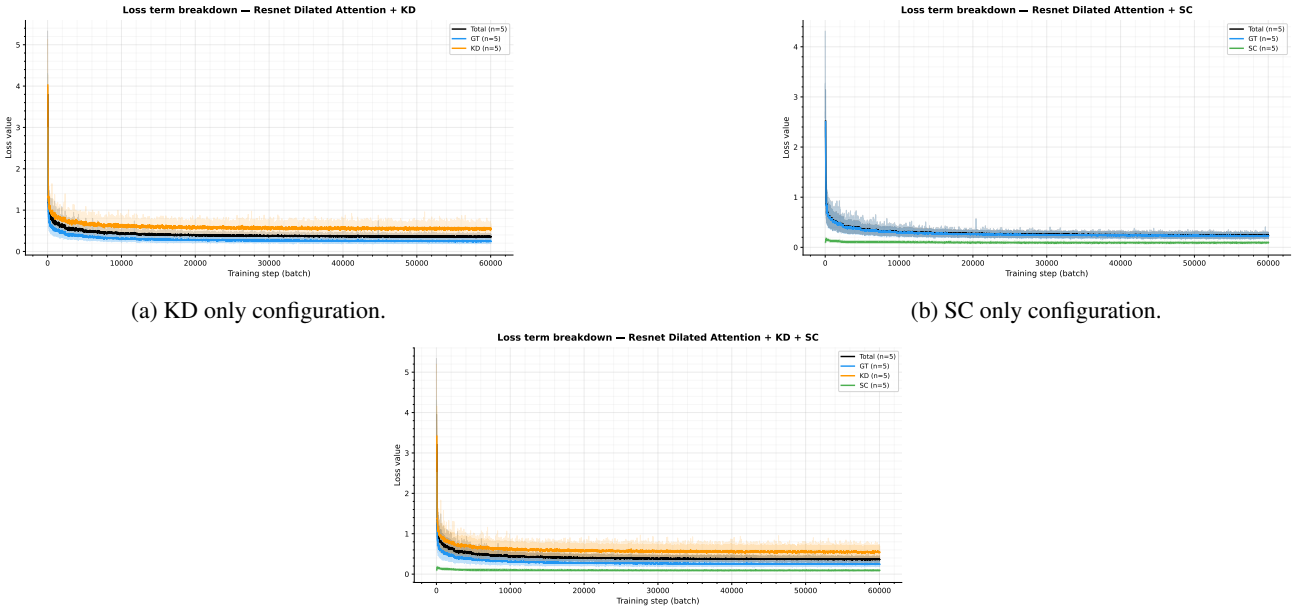
For (ii), relative improvements are computed on a per-seed basis. Specifically, for each pair of configurations and each random seed, we first compute the corresponding improvement and then report the empirical mean and standard deviation of these improvements across seeds. While the mean improvement can equivalently be obtained from the difference of the reported mean scores (by linearity of expectation), the same does not hold for the standard deviation. Since scores obtained from the same seed can generally be correlated across configurations, the variance of an improvement cannot be obtained from the independent score variances. Instead, we reporting the empirical distribution of paired per-seed improvements as an accurate characterization of both the mean of the improvement and its standard deviation.



(a) Mean± std training loss.

(b) Mean± std validation SILog-RMSE.

Figure 4. Train and validation loss curves during training across all epochs for all 5 seeds.



(a) KD only configuration.

(b) SC only configuration.

(c) KD + SC configuration.

Figure 5. Loss breakdown across all batches for different configurations.

Table 2. Pairwise improvement matrix — Test SILog-RMSE. Same convention as Table 1.

	VGG	RN18	RN18+D	RN18+D+A	+KD	+SC	+KD+SC
VGG	—	$+0.114_{\pm 0.019}$	$+0.124_{\pm 0.017}$	$+0.122_{\pm 0.015}$	$+0.139_{\pm 0.016}$	$+0.140_{\pm 0.016}$	$+0.149_{\pm 0.017}$
RN18	$-0.114_{\pm 0.019}$	—	$+0.010_{\pm 0.010}$	$+0.007_{\pm 0.007}$	$+0.025_{\pm 0.009}$	$+0.025_{\pm 0.013}$	$+0.034_{\pm 0.011}$
RN18+D	$-0.124_{\pm 0.017}$	$-0.010_{\pm 0.010}$	—	$-0.003_{\pm 0.005}$	$+0.015_{\pm 0.006}$	$+0.015_{\pm 0.006}$	$+0.024_{\pm 0.004}$
RN18+D+A	$-0.122_{\pm 0.015}$	$-0.007_{\pm 0.007}$	$+0.003_{\pm 0.005}$	—	$+0.017_{\pm 0.002}$	$+0.018_{\pm 0.006}$	$+0.027_{\pm 0.005}$
+KD	$-0.139_{\pm 0.016}$	$-0.025_{\pm 0.009}$	$-0.015_{\pm 0.006}$	$-0.017_{\pm 0.002}$	—	$+0.001_{\pm 0.005}$	$+0.010_{\pm 0.004}$
+SC	$-0.140_{\pm 0.016}$	$-0.025_{\pm 0.013}$	$-0.015_{\pm 0.006}$	$-0.018_{\pm 0.006}$	$-0.001_{\pm 0.005}$	—	$+0.009_{\pm 0.003}$
+KD+SC	$-0.149_{\pm 0.017}$	$-0.034_{\pm 0.011}$	$-0.024_{\pm 0.004}$	$-0.027_{\pm 0.005}$	$-0.010_{\pm 0.004}$	$-0.009_{\pm 0.003}$	—