2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

MaskingDepth: Masked Consistency Regularization for Semi-Supervised Monocular Depth Estimation

Jongbeom Baek[*,1], Gyeongnyeon Kim[*,1], Seonghoon Park[*,1], Honggyu An[3], Matteo Poggi[2], Seungryong Kim[†,3]

***Abstract*— We propose MaskingDepth, a semi-supervised learning framework for monocular depth estimation. MaskingDepth is designed to enforce consistency between the depths obtained from strongly-augmented images and the pseudo-depths derived from weakly-augmented images, which enables mitigating the reliance on large ground-truth depth quantities. In this framework, we leverage uncertainty estimation to only retain high-confident depth predictions from the weakly-augmented branch as pseudo-depths. We also present a novel data augmentation, dubbed K-way disjoint masking, that takes advantage of a naïve token masking strategy as an augmentation, while avoiding its scale ambiguity problem between depths from weakly- and strongly-augmented branches and risk of missing small-scale objects. Experiments on KITTI and NYU-Depth-v2 datasets demonstrate the effectiveness of each component, its robustness to the use of fewer depth-annotated images, and superior performance compared to other state-of-the-art semi-supervised learning methods for monocular depth estimation.**

I. Introduction

Monocular depth estimation, aiming to predict a dense depth map from a single image, has been one of the most essential tasks in numerous applications, such as robotics [1], augmented/virtual reality [2], or autonomous driving [3].

As a pioneering work, Eigen et al. [4] first introduced deep learning-based approach, and several following works [5], [6], [7], [8], [9], [10], [11], [12], [13] have achieved higher accuracy throughout the years. These methods were mostly formulated in a *supervised* learning regime, which requires a large number of images and corresponding ground-truth depths. However, these data are notoriously challenging to obtain [14], [15], compared to other types of annotation, such as image class labels [16] or segmentation labels [17]. To overcome this problem, *self-supervised* learning techniques [18], [19], [20], [21] have emerged, which formulate monocular depth estimation as an image reconstruction problem. Although this seems to be an attractive solution, these methods often require extra data, such as stereo pairs or video sequences which are not always available.In addition, they are known to often generate blurred depth at object boundaries [22], [23].

Some works [24], [25] have attempted to propose *semi-supervised* learning approaches by simply combining supervised learning and self-supervised learning [18], [19], [20], but they directly inherit limitations of existing self-supervised learning methods. Some works using stereo matching-based knowledge distillation [26], [27], [22] have also been proposed. However, they are constrained by the need for stereo image pairs and additional computation costs for training a stereo matching model.

(a) RGB images (b) Baseline (c) **MaskingDepth**

Fig. 1: **Effectiveness of our approach.** Our semi-supervised learner, dubbed **MaskingDepth**, produces high quality depth maps compared to supervised baseline by effectively leveraging a large amount of **unlabeled** data.

In this paper, for the first time, we present a novel semi-supervised learning framework for monocular depth estimation, called **MaskingDepth**, based on an uncertainty-aware consistency regularization. Our framework enforces consistency between depths obtained from strongly-augmented images and pseudo-labels obtained from weakly-augmented images, while the uncertainty estimation technique [28], [29] aids depth consistency and facilitates convergence by filtering out the noise of pseudo-labels.

To apply perturbations to an input image in the consistency regularization, we propose a new data augmentation, called K-way disjoint masking, inspired by token masking strategies for Transformers [30], [31], [32]. Although the naïve masking technique [30], [31], [32] yielded superior performance on classification tasks such as image classification and semantic segmentation, adopting this to monocular depth estimation as data augmentation, which heavily relies on context information, may cause scale ambiguity and omit the context of small objects [32]. To overcome this, the K-way disjoint masking jointly decodes scattered tokens that are encoded from a K-disjoint set of tokens independently, and thus mitigates the scale ambiguity and restores the full context from the image while introducing masked interaction in self-attention of encoders as a perturbation during training. In our framework, we encourage depth and feature consis-

[*]Equal contribution [†]Corresponding author
[1]Korea University, Korea [2]University of Bologna, Italy
[3]Korea Advanced Institute of Science and Technology, Korea

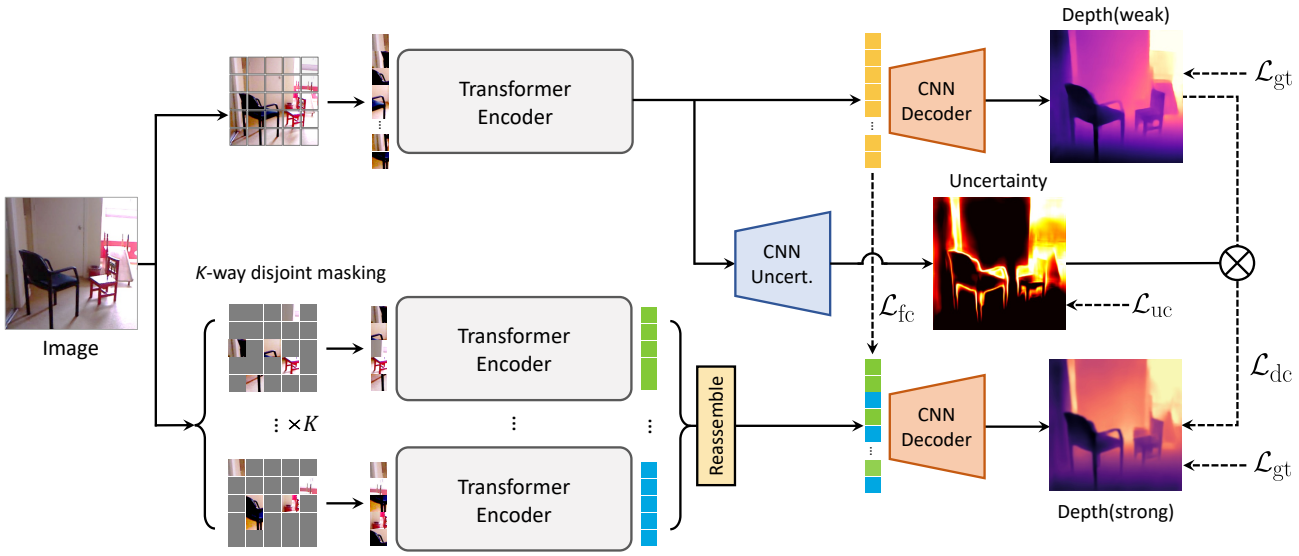


Fig. 2: **Overview of MaskingDepth architecture.** It consists of two main components; a branch using full tokens (top), and a branch using K-way disjoint masked tokens (bottom), where K-number of subsets are encoded independently and concatenated before decoding. We use consistency loss $\mathcal{L}_{\text{dc}}$ to make predictions between original and augmented images consistent, aided by an uncertainty measure. Feature consistency loss $\mathcal{L}_{\text{fc}}$ is also applied to facilitate the convergence.

tencies across two branches from two augmented views by K-way disjoint masking.

In the experiments, we evaluate MaskingDepth on standard benchmarks, including KITTI [14] and NYU-Depth-v2 [15], showing outstanding performance compared to previous methods. We validate each component through an extensive ablation study.

II. Related Work

A. Monocular Depth Estimation

Monocular depth estimation aims at estimating a depth from a single image. [4] first tackled this task with deep neural networks. Since then, several approaches [5], [6], [7], [8], [9], [10], [11], [12], [13] based on supervised learning have been presented to improve performance. Although these methods have achieved remarkable accuracy over traditional, handcrafted approaches [33], [34], their success depends on massive amounts of ground-truth depth maps that require a labor-intensive process for collection and cleaning [14], [15].

To address this limitation, self-supervised learning methods [35], [19], [20] formulate image reconstruction problem, leveraging geometric information. They have emphasized the importance to mitigate the dependency on annotations, but often produces indistinct depth result near object boundaries [22], [23]. Unlike both of the aforementioned approaches, there has not been much work on semi-supervised depth estimation. [24], [25] simply combined supervised and self-supervised loss functions. Recently, several works [36], [27], [37], [38] have attempted to distill stereo knowledge for monocular depth estimation. However, they are still constrained by the need for specific data (stereo pairs) and additional computation cost (training a stereo module). Our framework alleviates the reliance on labeled data by leveraging consistency regularization to use unlabeled data.

B. Token Masking

Token masking has been widely used to learn representations by reconstructing images that are corrupted by masking [30], [31], [32]. After BERT [39] proposed the masked language modeling task, one of the most successful methods for pre-training in NLP, related works have explored a variety of masked image prediction strategies suited for Transformers [30], [32]. ViT [40] studied a masked patch prediction to facilitate representation learning, and BEiT [30] extended upon this by predicting discrete tokens. Recent literature [31], [32] introduce an extremely simple yet effective approach. Masked autoencoder (MAE) [32] utilizes only unmasked tokens to encode meaningful representations. In addition, MRA [41] leverages this strategy to generate augmented images. In this paper, we propose an effective data augmentation strategy that leverages token masking.

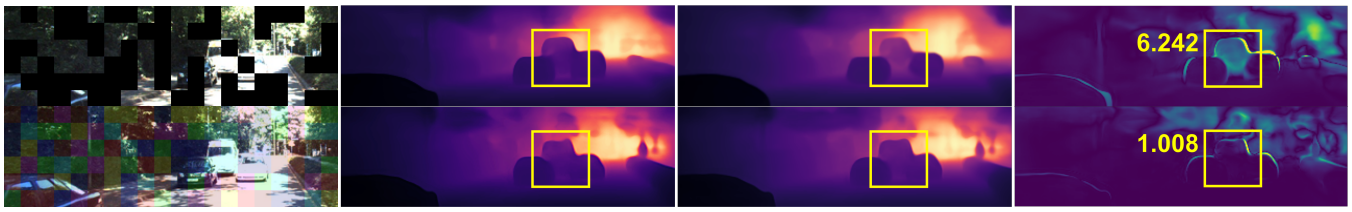
III. Methodology

A. Problem Formulation

Let us denote a color image and its corresponding depth map as I and D, respectively. The objective of monocular depth estimation is to learn a mapping function $f(\cdot)$ from the image I to its corresponding depth D such that $D = f(I)$. Recent learning-based methods formulate the mapping function with convolutions [4], [6] or Transformers [42], [43] as a neural network f_θ with parameters θ. To train the monocular depth estimation networks f_θ in a supervised manner, the ground-truth D_{gt} is required, but building large-scale dense depth data is notoriously challenging [35], [44]. In addition, to alleviate depth capture errors, post-processing [45], [25] is essential, which introduces further burden.

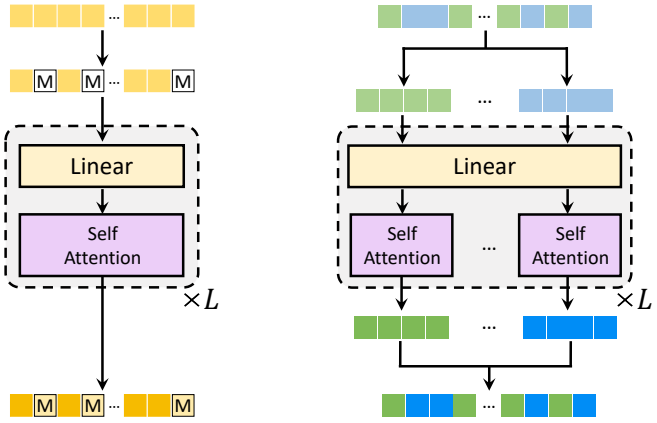
B. Motivation and Overview

In this paper, we present a novel semi-supervised learning framework that facilitates the model to learn monocular depth



(a) Masked RGB images (b) Depths (weak) (c) Depths (strong) (d) Difference maps

Fig. 3: **Effectiveness of our masking strategy to handle scale ambiguity:** In (a), naïve masking is applied above and our masking is applied below. Our masking shows dividing independent subsets. The first row of (b) and (c) are naïve masking results, and the second row of (b) and (c) are our masking results. In (d), we visualize the difference map between (b) and (c). We denote the mean scale difference in the boxed area. Our method better generates scale-consistent results, which in turn helps to better learn the monocular depth estimation networks.



(a) Naïve token masking (b) K-way disjoint masking

Fig. 4: **Illustration of masking:** (a) naïve token masking [31], [30] and (b) our K-way disjoint masking. Compared to naïve masking, K-way masking consistently well produces dense depth maps because it considers the entire set of tokens during decoding, while providing sufficient perturbations to the inputs.

estimation on a large number of unlabeled data with limited depth annotations by introducing consistency regularization between two differently augmented views from the same image. Compared to other semi-supervised learning methods [24], [36], [27], [37], [38], [25] that require the dataset consisting of stereo pairs or a sequence of frames for training a stereo network or reconstructing an image, our approach does not have the dependency problem.

Given an image I, we build two different branches, one for processing an image passed through weak augmentation (called weak branch) and the other for a strongly augmented image (called strong branch), where the consistency between the two images through the networks f_θ is encouraged. In particular, a weakly-augmented image I_{weak} and a strongly-augmented image I_{strong} are fed to the network f_θ, and then consistency is defined as follows:

$$\mathcal{L} = \mathcal{D}(\text{sg}(f_\theta(I_{\text{weak}})), f_\theta(I_{\text{strong}})), \tag{1}$$

where $\text{sg}(\cdot)$ is a stop-gradient operation [46] and $\mathcal{D}(\cdot,\cdot)$ is a distance function like mean squared error (MSE) [47]. In this setting, we can interpret $f_\theta(I_{\text{weak}})$ as a pseudo-label. To effectively implement this strategy, appropriate data augmentation techniques are important.

However, it is challenging to make difference between two branches since depth-specific data augmentation techniques have been rarely studied in monocular depth estimation [48], [49]. Furthermore, conventional data augmentation techniques such as crop [50] and rotation [51], effective in classification, are no longer effective for monocular depth estimation as they can lead to geometric inconsistency [52] between the two branches.

To address this issue, we present a novel data augmentation technique, inspired by token masking [31], [32], which allows for generating geometrically consistent depth maps while applying sufficient perturbations to the inputs. As illustrated in Fig. 2, our framework follows the backbone model f_θ by consisting of a Transformer-based encoder f_θ^{enc}, which takes a tokenized image and outputs encoded features, and a CNN-based decoder f_θ^{dec} [13]. In addition, we encourage not only feature similarities [30] but also depth similarities between the two branches processing the two augmented views. To aid the latter, we present uncertainty estimation [28], [29] that helps the convergence of training by filtering out the noise of pseudo labels.

C. Naïve Token Masking and Its Limitations

Recent token masking techniques for Transformers [30], [31], [32], [41] have shown their effectiveness as data augmentation. The most naïve way to formulate token masking techniques as augmentation is to simply mask out the tokens. Specifically, given an image $I \in \mathbb{R}^{h\times w\times 3}$, we reshape it into a sequence of flattened non-overlapped 2D patches $X \in \mathbb{R}^{N\times P}$, where $h \times w$ is the resolution of the original image, $P = p \times p \times 3$ and $p \times p$ is the resolution of each image patch, and $N = hw/p^2$. These flattened 2D patches X are embedded by a trainable linear projection [40] operator, which proceeds to be fed into the Transformer encoder f_θ^{enc}. By applying the randomly sampled mask, the sequence of flattened 2D patches X can be transformed into X'. Also, similar techniques were used to increase the robustness of Transformers for image-level or pixel-level classification [31], [32].

As shown in Fig. 3, applying the naïve masking strategy to monocular depth estimation causes scale ambiguity issue. Since this task inherently has scale ambiguity problem, there are multiple scale values to construct depth maps on the missing region while keeping coherence with the given

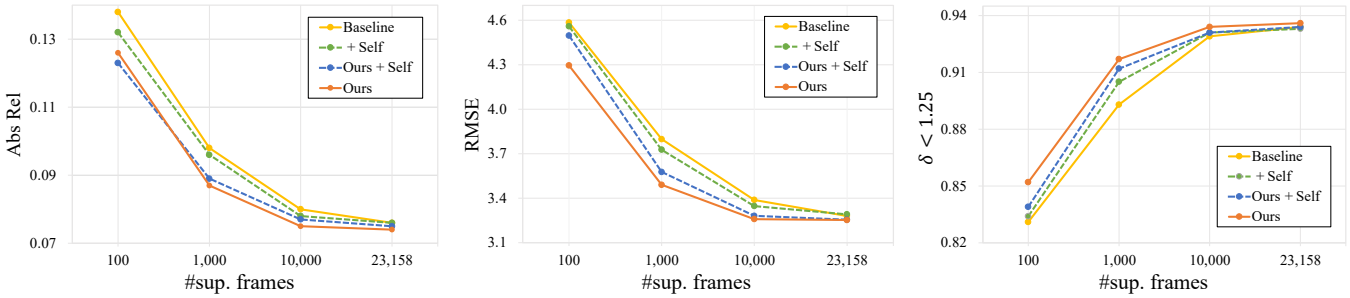


Fig. 5: **Quantitative results on the KITTI dataset in a sparsely-supervised setting.** 'Baseline' only uses a sparse depth, and 'Self' indicates existing self-supervised strategies [35], [20]. 'Ours' indicates the proposed semi-supervised framework.

context. Also, there is the risk of missing out small-scale objects due to heavily dependency on contextual information.

D. K-way Disjoint Masking

To overcome the limitations of naïve masking, we present a K-way disjoint masking technique, where a K-disjoint set of tokens are encoded independently, then concatenated and decoded simultaneously, as illustrated in Fig. 4. Similar to naïve masking, our K-way disjoint masking introduces masked interaction in self-attention of encoders as a perturbation during training. But, by capturing the entire scene from the partially divided inputs, our method can reduce inherent ambiguity [53] and lead to coherent results, as shown in Fig. 3. Moreover, since it ensures scale consistency by keeping the image size and orientation unaltered, our method can act as data augmentation.

Specifically, we divide the sequence of flattened 2D patches $X \in \mathbb{R}^{N\times P}$ into K non-overlapping subsets X_k for $k \in \{1, ..., K\}$, with $X_k \in \mathbb{R}^{M\times P}$, where M is set to be a random value smaller than N to avoid learning with a fixed size of the tokens set. In other words, the concatenation of all X_k tokens should reconstruct the original token representation X, while maintaining the proper position ordering such that

$$X = [X_1, X_2, ..., X_K], \tag{2}$$

where $[\cdot, \cdot]$ denotes a concatenation operator. By independently encoding X_k to the latent vector $\mathbf{z}_k$ such that $\mathbf{z}_k = f_\theta^{\text{enc}}(X_k)$, unlike original Transformer-based encoding, i.e., $\mathbf{z} = f_\theta^{\text{enc}}(X)$, the limited attention candidates are considered when running self-attention computation, which in turn implements an augmentation over tokens.

Then, to decode all the $\mathbf{z}_k$, we reassemble them as $\bar{\mathbf{z}} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_K]$ and obtain the final depth $D = f_\theta^{\text{dec}}(\bar{\mathbf{z}})$.

In our framework, we control the intensity of augmentations to the networks by adjusting K. In our experiments, we empirically set $K = 1$ for the weak branch, and $K = 64$ for the strong branch, which generate decoded depth maps D_{weak} and D_{strong}, respectively.

E. Loss Functions

To train the networks, we adopt a sparse supervised loss and the proposed unsupervised loss. In addition, we adopt a loss for modeling the uncertainty of the pseudo ground-truth produced by the weak branch and a feature consistency loss. $\|\cdot\|_1$ and $\|\cdot\|_2$ are ℓ_1 and ℓ_2 loss functions, respectively.

Supervised loss. When sparsely depth-labeled data is available to train the network, we can minimize the supervised loss function $\mathcal{L}_{\text{gt}}$ between the predicted D and sparse ground-truth D_{gt} such that

$$\mathcal{L}_{\text{gt}} = \|D - D_{\text{gt}}\|_1. \tag{3}$$

A small number of fully annotated data used in a supervised manner across both branches allows the network model to ignite the learning process, which is then carried out mainly on unlabeled data through consistency regularization.

Depth consistency loss. Our loss function encourages depth prediction of the strongly-augmented image to be close to the prediction of the weakly-augmented image, enabling pixel-level learning without the need for annotated ground-truths, thus it serving as an effective solution to data hunger caused by sparse annotations. The depth consistency loss $\mathcal{L}_{\text{dc}}$ assisted by the uncertainty map $U(D_{\text{weak}})$ can be written as:

$$\mathcal{L}_{\text{dc}} = \frac{\|\text{sg}(D_{\text{weak}}) - D_{\text{strong}}\|_1}{\text{sg}(U(D_{\text{weak}}))}. \tag{4}$$

where $U(D)$ denotes the uncertainty map related to the predicted depth map D.

Uncertainty loss. To filter out the noise on pseudo labels, we train the uncertainty module. This module, which is a key ingredient in our framework, allows for transferring only reliable depth knowledge from weak branch to strongly augmented branch. To model such uncertainty, we leverage a negative log-likelihood minimization [28] as:

$$\mathcal{L}_{\text{uc}} = \frac{\|D_{\text{weak}} - D_{\text{gt}}\|_1}{U(D_{\text{weak}})} + \log(U(D_{\text{weak}})), \tag{5}$$

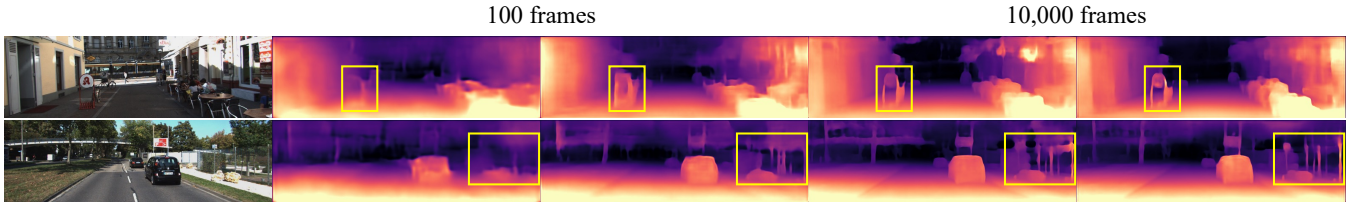
By training the network to model its uncertainty, predictions on unlabeled data will be trusted if highly confident.

Feature consistency loss. Within our framework, geometric distortions are not applied to the two branches, thus the encoded feature consistency can also be encouraged. The feature consistency loss is then defined as

$$\mathcal{L}_{\text{fc}} = \|\mathbf{z}_{\text{weak}} - h(\mathbf{z}_{\text{strong}})\|_2, \tag{6}$$

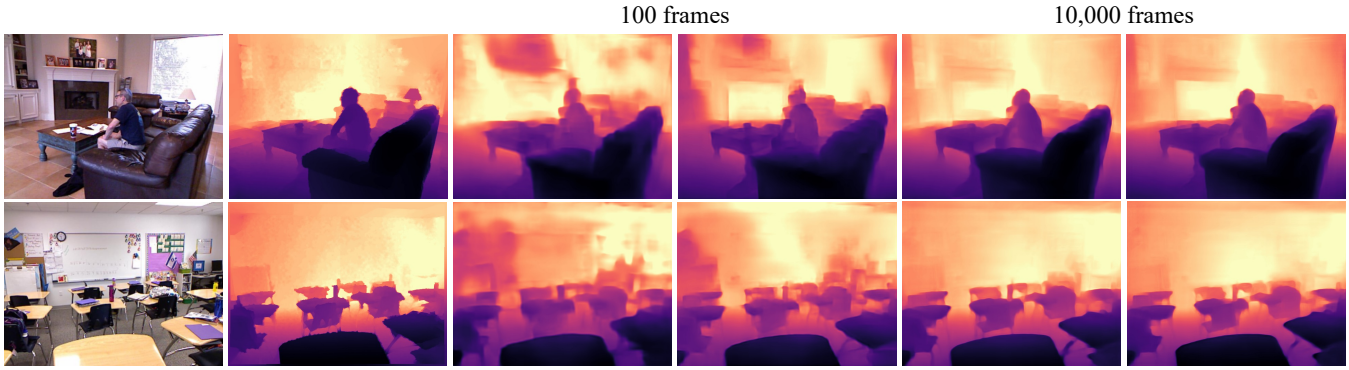
where $h(\cdot)$ is the additional MLP predictor head, which provides better results as shown in the literature [46], [54] and prevents collapse [55].

Total loss. By considering all the loss terms, the final loss is defined as $\mathcal{L} = \mathcal{L}_{\text{gt}} + \lambda_{\text{dc}}\mathcal{L}_{\text{dc}} + \lambda_{\text{uc}}\mathcal{L}_{\text{uc}} + \lambda_{\text{fc}}\mathcal{L}_{\text{fc}}$, where λ_{dc}, λ_{uc}, and λ_{fc} represent hyper-parameters.



(a) RGB images (b) Baseline (c) Ours (d) Baseline (e) Ours

Fig. 6: **Qualitative results on the KITTI dataset [14]:** (a) RGB images, predicted depth maps by (b), (d) baseline, and (c), (e) ours using 100 and 10,000 labeled frames, respectively.



(a) RGB images (b) GT depths (c) Baseline (d) Ours (e) Baseline (f) Ours

Fig. 7: **Qualitative results on the NYU-Depth-v2 dataset [15]:** (a) RGB images, (b) ground-truth depth maps, and predicted depth maps by (c), (e) baseline, and (d), (f) ours using 100 and 10,000 labeled frames, respectively.

IV. Experiments

A. Implementation Details

We implemented our **MaskingDepth** with the PyTorch library [56]. We conduct all our experiments on 24GB RTX-3090 GPUs, using DPT-Base [13] as a backbone model. We set the learning rate to 10^{-5} for the encoder and 10^{-4} for the decoder. The encoder is initialized with ImageNet-pretrained [57] weights, whereas the decoder is initialized randomly. We train the entire model with batch size 8 and use Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

To avoid collapsing, we balance the ratio of labeled and unlabeled data in one batch to 1:7 similarly to [16]. Besides our new data augmentation approach, we adopt flipping and jittering, widely used in the literature [58], [59]. For confidence estimation, we train the network to predict the log variance because it is more numerically stable [28] than regressing the variance itself, as the loss avoids any division by zero. We use an identical hyperparameter set (i.e., $\lambda_{\text{uc}} = 1$, $\lambda_{\text{dc}} = 1$, $\lambda_{\text{fc}} = 1$, $K = 64$ for strong augmentation) for all experiments unless otherwise specified.

B. Experimental Settings

Datasets. We first evaluate the performance of MaskingDepth and others on the KITTI dataset [14] and NYU-Depth-v2 [15]. The KITTI dataset [14] provides outdoor scenes captured by 3D laser data. The RGB images are resized to 640×192 resolutions for training. We follow the standard Eigen training/testing split [4]. We use randomly sampled 10,000, 1,000, and 100 images from $24K$ (i.e., left frames in the Eigen training split) for labeled images during training. We evaluate our trained model on 652 annotated test images for single-view depth estimation, using the improved ground truth by [45]. The NYU-Depth-v2 dataset [15] is composed of various indoor scenes and corresponding depth maps at a resolution of 640×480. We train our network on the same number of labeled images as KITTI, randomly sampled from the original $40K$ total. The remaining images in the training set are used as unlabeled images. We test our trained model on 654 test images.

Evaluation metrics. In our experiments, we follow the standard evaluation protocol of the prior work [4] to evaluate the effectiveness of MaskingDepth. The error metrics are defined as Absolute Relative error (AbsRel), Squared Relative error (SqRel), Root Mean Squared Error (RMSE), Root Mean Squared log Error (RMSElog), and accuracy under the threshold (< 1.25) (δ).

C. Experimental Results

In this section, we investigate the effects of sparse labels on supervised depth training, and how MaskingDepth is able to mitigate the degradation of depth maps when the number of labels is significantly limited. Especially, our method is useful for improving performance by utilizing vast unlabeled data in place of expensive ground-truth depth annotations.

Robustness of MaskingDepth. As a baseline, we compare DPT-Base [13] trained on supervised and conventional semi-supervised manners obtained with self-supervised losses [35], [20]. Results for the KITTI dataset [14] using different numbers of supervised frames are shown in Fig. 5. MaskingDepth outperforms the baseline through consistency regularization using unlabeled frames

As the amount is further decreased, the performance of all approaches, except for ours, shows a significant and rapid

Method	Sup.	Self-Sup.			Data Setting		AbsRel↓	SqRel↓	RMSE↓	RMSElog↓	δ↑
		Video	Stereo	Cons.	label	unlabel					
DORN [10]	✓	-	-	-	K	-	0.072	0.307	2.727	0.120	0.932
Yin et al. [60]	✓	-	-	-	K	-	0.072	-	3.258	0.117	0.938
DPT-Hybrid [13]	✓	-	-	-	K+Mix	-	**0.062**	**0.222**	**2.575**	**0.092**	**0.959**
DPT-Base (Baseline)	✓	-	-	-	K	-	0.071	0.292	2.964	0.107	0.942
DPT-Hybrid (Baseline)	✓	-	-	-	K	-	0.068	0.251	2.717	0.099	0.954
Monodepth2* [20]	-	✓	✓	-	-	K	0.080	0.466	3.681	0.127	0.926
ManyDepth* [21]	-	✓	-	-	-	K	**0.070**	**0.399**	**3.455**	**0.113**	**0.941**
SVSM FT [61]	✓	-	✓	-	K+F	K+F	0.077	0.392	3.569	0.127	0.919
Kuznietsov et al. [24]	✓	-	✓	-	K	K	0.089	0.478	3.610	0.138	0.906
Baek et al. [23]	✓	✓	✓	-	K	K	0.078	0.381	3.404	0.121	0.930
Amiri et al. [25]	✓	-	✓	-	K	K+C	0.078	0.417	3.464	0.126	0.923
MaskingDepth (DPT-Base)	✓	-	-	✓	K	C	0.067	0.285	2.932	0.104	0.947
MaskingDepth (DPT-Hybrid)	✓	-	-	✓	K	C	**0.063**	**0.235**	**2.653**	**0.095**	**0.958**

TABLE I: **Quantitative results on the Eigen split of the KITTI dataset [14].** 'Sup.' and 'Self-Sup. (Video and Stereo)' indicate supervised, existing self-supervised strategies on video and stereo pairs, respectively. 'Self-Sup. (Cons.)' denotes our proposed consistency regularization, which needs **no stereo images or video sequences** '$*$' means calibrated scale results by using the per-image median ground truth scaling [18]. 'K', 'C', and 'F' indicate KITTI [14], Cityscapes [62], and FlyingThings3D [61], respectively. 'Mix' indicates the dataset proposed from [13], containing $1.4M$ images, which is approximately 60 times larger than KITTI.

Method	Data Setting		AbsRel↓	RMSE↓	δ↑
	label	unlabel			
DORN	N	-	0.115	0.509	0.828
BTS	N	-	0.110	0.392	0.885
DPT-Hybrid	N+Mix	-	0.110	**0.357**	**0.904**
DPT-Base (Baseline)	N	-	0.106	0.380	0.899
MaskingDepth (DPT-Base)	N	S	**0.104**	0.372	**0.904**

TABLE II: **Quantitative results on the NYU-Depth-v2 [15].** 'N' and 'S' indicate NYU-Depth-v2 and SUN RGB-D [63] datasets, respectively.

Methods	L2 distance↓	Cosine similarity↑
Baseline	1.129 ± 0.006	0.643 ± 0.001
Naïve masking	0.986 ± 0.008	0.743 ± 0.003
Ours	$\mathbf{0.625 \pm 0.002}$	$\mathbf{0.806 \pm 0.001}$

TABLE III: **Effectiveness of encoded features.**

Methods	RMSE↓	Std of RMSE ↓	Max of RMSE ↓
Naïve masking	3.44 ± 0.05	3.20 ± 0.14	28.29 ± 1.60
Ours	$\mathbf{2.17 \pm 0.02}$	$\mathbf{1.91 \pm 0.05}$	$\mathbf{16.43 \pm 1.01}$

TABLE IV: **Instance-wise scale error on KITTI [14].**

decline. This is mostly due to the model's inability to learn the appropriate scale and structure of a scene with such sparse information. However, as the labels become sparser, the performance degradation of our proposed method progresses more slowly compared to others(baseline and naive semi-supervised approach using self-supervised losses), and the performance gap gets larger. Note that when MaskingDepth is incorporated with the existing self-supervised approach, the performance gain was marginal because the self-supervised loss function increase inherent scale ambiguity [64]. We also provide a qualitative comparison of the baseline and our method on the KITTI dataset in Fig. 6 and the NYU-Depth-v2 dataset in Fig. 7.

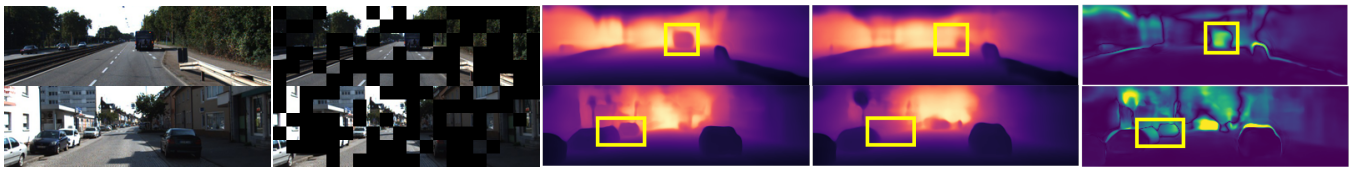
Comparison to other methods. As our method does not rely on stereo or video sequence frames, it is agnostic to the configuration of the unlabeled training set. Table I compares our semi-supervised method that uses additional data against existing approaches. We trained our model on the KITTI dataset [14] as labeled data and the Cityscape dataset [62] as additional unlabeled data. Our approach used both supervised loss and uncertainty loss functions for the labeled data(KITTI), whereas only a consistency loss was applied to unlabeled data(Cityscape). In this experiment, we follow image resolution [13]. In the results, our method achieves significant improvement in comparison to the baseline by utilizing unlabeled data and surpasses the state-of-the-art in semi-supervised depth estimation methods. Moreover, despite using a smaller model capacity and a fewer annotated data, MaskingDepth shows competitive performance against DPT-Hybrid. A similar trend can be seen in Table II, where our method utilizes the SUN RGB-D [63] dataset as additional unlabeled data and NYU-Depth-v2 as labeled data.

D. Ablation Study

We analyze the effectiveness of different design choices in our framework on the KITTI dataset [14]. We exploit 10,000 and 100 randomly sampled annotated images, respectively for masking strategy and other ablation studies.

Masking strategy. In this section, we extensively analyze our K-way disjoint masking to prove the effectiveness of our masking strategy. We evaluate feature similarity between global interaction and masked interaction in Table III. Our masking makes the masked interaction follow the global interaction well and helps to learn good representation while using naïve masking limitedly learns representation. In addition, to demonstrate that naïve masking [31], [32] definitely leads to scale ambiguity in monocular depth estimation, we evaluate instance-wise scale error in Table IV. Fig. 8 shows missing objects, even when instances are not entirely masked. Since our augmentation captures the entire scene, it reduces the inherent scale ambiguity compared to naïve masking.

Loss functions. We examine each component of the loss function in our method. It consists of four components:



(a) RGB images (b) Masked RGB images (c) Depths (weak) (d) Depths (strong) (e) Difference maps

Fig. 8: **Missing object cases of naïve masking on the KITTI dataset [14]:** (a) RGB images, (b) masked RGB images, (c) depth maps predicted from (a), (d) depth maps predicted from (b), and (e) difference maps between (c) and (d).

Methods	D	U	F	AbsRel ↓	RMSE ↓	δ↑
Baseline	-	-	-	0.136	4.585	0.833
Ours	✓	-	-	0.132	4.355	0.848
	✓	✓	-	0.126	4.296	0.851
	-	-	✓	0.131	4.422	0.849
	✓	✓	✓	**0.124**	**4.263**	**0.855**

TABLE V: **Ablation study on main components:** Depth consistency (D), uncertainty (U), and feature consistency (F).

the sparse supervised term, the depth consistency term, the feature consistency term, and the uncertainty term. To evaluate the impact of each component, we start by using only the sparse supervised loss as a baseline and then study the effect of adding each of the other three components. As shown in Table V, the performance of our network improves as each component is added. We can see that using all components together leads to a significant improvement.

The number of K. To study the influence of the masking ratio, we train the network by adopting different values of K for the strong branch, respectively $K = 4, 16, 64$, and 128. Table VI shows the quantitative evaluation results for this study. Starting from $K = 4$, the error decreases with the increase of K, until degrading for $K = 128$. We set $K = 64$ as the default since it yields the best results.

Predictor head. To improve representation power of the encoder, we consider feature consistency loss between encoded features. When the predictor head was removed, collapsing occurred and training did not proceed. Our framework without predictor head is conceptually analogous to naive Siamese network, which could not avoid collapsing [55]. In this set of experiments we evaluate the performance of the predictor head used for providing better results in feature consistency loss. Results are shown in Table VII. Although one block of Transformer and MLP showed comparable performance, a simple MLP layer is much more efficient.

V. Conclusion

In this paper, we presented **MaskingDepth**, a novel semi-supervised framework for monocular depth estimation using consistency regularization. MaskingDepth leverages depth-unlabeled images without requiring stereo or sequential frames. We proposed a data augmentation method, K-way disjoint masking, which produces scale-consistent depth maps and stabilizes the consistency regularization framework. Additionally, uncertainty estimation effectively mitigates performance degradation by filtering noise on pseudo labels. Our method showed significant improvement on extremely sparse labeled data and outperforms other semi-supervised approaches.

K	AbsRel ↓	RMSE ↓	δ↑
4	0.131	4.456	0.850
16	0.128	4.298	**0.855**
64	**0.124**	**4.263**	**0.855**
128	0.132	4.382	0.849

TABLE VI: **Influence of the number of K.**

Method	Blocks	AbsRel ↓	δ↑
w/o head	-	0.317	0.423
Transformer	1	0.125	**0.856**
Transformer	2	0.129	0.848
MLP	2	**0.124**	0.855

TABLE VII: **Comparison of different predictor heads.**

Acknowledgement

This research was supported by the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00348469, RS-2023-00266509), Autonomous Driving Center, R&D Division, Hyundai Motor Company and National Research Foundation of Korea (RS-2024-00346597).

References

[1] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1281–1292.

[2] W. Lee, N. Park, and W. Woo, "Depth-assisted real-time 3d object detection for augmented reality," in *ICAT*, vol. 11, no. 2, 2011, pp. 126–132.

[3] V. R. Kumar, S. A. Hiremath, M. Bach, S. Milz, C. Witt, C. Pinard, S. Yogamani, and P. Mäder, "Fisheyedistancenet: Self-supervised scale-aware distance estimation using monocular fisheye camera for autonomous driving," in *ICRA*. IEEE, 2020, pp. 574–581.

[4] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *NIPS*, vol. 27, 2014.

[5] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *CVPR*, 2015, pp. 1119–1127.

[6] S. Kim, K. Park, K. Sohn, and S. Lin, "Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields," in *ECCV*, 2016, pp. 143–159.

[7] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in *CVPR*, 2017, pp. 5038–5047.

[8] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *PAMI*, vol. 38, no. 10, pp. 2024–2039, 2015.

[9] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[10] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *CVPR*, 2018, pp. 2002–2011.

[11] 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*, 2019.

[12] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *PAMI*, 2020.

[13] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *ICCV*, 2021, pp. 12 179–12 188.

[14] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*. IEEE, 2012, pp. 3354–3361.

[15] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*. Springer, 2012, pp. 746–760.

[16] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *NIPS*, vol. 33, pp. 596–608, 2020.

[17] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, "Pseudoseg: Designing pseudo labels for semantic segmentation," *arXiv*, 2020.

[18] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017, pp. 1851–1858.

[19] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017, pp. 270–279.

[20] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019, pp. 3828–3838.

[21] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1164–1174.

[22] H. Choi, H. Lee, S. Kim, S. Kim, S. Kim, K. Sohn, and D. Min, "Adaptive confidence thresholding for monocular depth estimation," in *ICCV*, 2021, pp. 12 808–12 818.

[23] J. Baek, G. Kim, and S. Kim, "Semi-supervised learning with mutual distillation for monocular depth estimation," in *ICRA*. IEEE, 2022, pp. 4562–4569.

[24] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *CVPR*, 2017, pp. 6647–6655.

[25] A. J. Amiri, S. Y. Loo, and H. Zhang, "Semi-supervised monocular depth estimation with left-right consistency using deep neural network," in *ROBIO*, 2019, pp. 602–607.

[26] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano, "Unsupervised domain adaptation for depth prediction from images," *PAMI*, vol. 42, no. 10, pp. 2396–2409, 2019.

[27] J. Cho, D. Min, Y. Kim, and K. Sohn, "A large rgb-d dataset for semi-supervised monocular depth estimation," *arXiv*, 2019.

[28] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *NIPS*, vol. 30, 2017.

[29] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *CVPR*, 2020, pp. 3227–3237.

[30] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv*, 2021.

[31] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," *arXiv*, 2021.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *arXiv*, 2021.

[33] A. Saxena, S. H. Chung, and A. Y. Ng, "3-d depth reconstruction from a single still image," *IJCV*, vol. 76, no. 1, pp. 53–69, 2008.

[34] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *CVPR*, 2014, pp. 89–96.

[35] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *ECCV*. Springer, 2016, pp. 740–756.

[36] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *ECCV*, 2018, pp. 484–500.

[37] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *CVPR*, 2019, pp. 9799–9809.

[38] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *ICCV*, 2019, pp. 2162–2171.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2018.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv*, 2020.

[41] H. Xu, S. Ding, X. Zhang, H. Xiong, and Q. Tian, "Masked autoencoders are robust data augmentors," *arXiv preprint arXiv:2206.04846*, 2022.

[42] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Transformer-based attention networks for continuous pixel-wise prediction," in *ICCV*, 2021, pp. 16 269–16 279.

[43] Z. Li, Z. Chen, X. Liu, and J. Jiang, "Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation," *arXiv*, 2022.

[44] J. Xie, R. Girshick, and A. Farhadi, "Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks," in *ECCV*. Springer, 2016, pp. 842–857.

[45] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *3DV*. IEEE, 2017, pp. 11–20.

[46] X. Chen and K. He, "Exploring simple siamese representation learning," in *CVPR*, 2021, pp. 15 750–15 758.

[47] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[48] Y. Ishii and T. Yamashita, "Cutdepth: Edge-aware data augmentation in depth estimation," *arXiv*, 2021.

[49] D. Kim, W. Ga, P. Ahn, D. Joo, S. Chun, and J. Kim, "Global-local path networks for monocular depth estimation with vertical cutdepth," *arXiv*, 2022.

[50] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv*, 2017.

[51] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv*, 2018.

[52] T. v. Dijk and G. d. Croon, "How do neural networks see depth in single images?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2183–2191.

[53] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016, pp. 2536–2544.

[54] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *NIPS*, vol. 33, pp. 21 271–21 284, 2020.

[55] C. Zhang, K. Zhang, C. Zhang, T. X. Pham, C. D. Yoo, and I. S. Kweon, "How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning," *arXiv*, 2022.

[56] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009, pp. 248–255.

[58] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *ICCV*, 2017, pp. 1301–1310.

[59] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *CVPR*, 2021, pp. 2918–2928.

[60] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5684–5693.

[61] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016, pp. 4040–4048.

[62] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016, pp. 3213–3223.

[63] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[64] R. McCraith, L. Neumann, and A. Vedaldi, "Calibrating self-supervised monocular depth estimation," *arXiv preprint arXiv:2009.07714*, 2020.