Learning Depth from Past Selves: Self-Evolution Contrast for Robust Depth Estimation

Jing Cao[1,2], Kui Jiang[1,2,*], Shenyi Li[1], Xiaocheng Feng[1], Yong Huang[3]

[1]Faculty of Computing, Harbin Institute of Technology,
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ),
[3]Key Lab of Smart Prevention and Mitigation of Civil Engineering Disasters, Harbin Institute of Technology
{caojing, 2022111995}@stu.hit.edu.cn, {jiangkui, huangyong}@hit.edu.cn, xcfeng@ir.hit.edu.cn

Abstract

Self-supervised depth estimation has gained significant attention in autonomous driving and robotics. However, existing methods exhibit substantial performance degradation under adverse weather conditions such as rain and fog, where reduced visibility critically impairs depth prediction. To address this issue, we propose a novel self-evolution contrastive learning framework called SEC-Depth for self-supervised robust depth estimation tasks. Our approach leverages intermediate parameters generated during training to construct temporally evolving latency models. Using these, we design a self-evolution contrastive scheme to mitigate performance loss under challenging conditions. Concretely, we first design a dynamic update strategy of latency models for the depth estimation task to capture optimization states across training stages. To effectively leverage latency models, we introduce a Self-Evolution Contrastive Loss (SECL) that treats outputs from historical latency models as negative samples. This mechanism adaptively adjusts learning objectives while implicitly sensing weather degradation severity, reducing the need for manual intervention. Experiments show that our method integrates seamlessly into diverse baseline models and significantly enhances robustness in zero-shot evaluations.

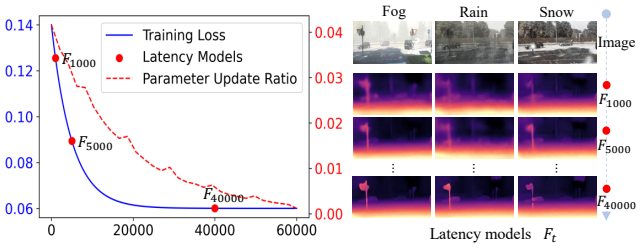

Figure 1: Illustration of latency models evolution. The left figure shows the relationship between training step t, training loss, and parameter update ratio, where a decreasing update ratio indicates model convergence. Models at different optimization steps t within the parameter space are defined as latency models F_t. The right figure presents the depth outputs of latency models under adverse weather conditions. We leverage these evolving latency models to construct negative samples for the contrastive learning, which encourages the depth model to learn robust representations from its own historical information.

Introduction

Accurate depth estimation from images is a critical computer vision task with significant applications in 3D scene reconstruction (Yin et al. 2022) and autonomous driving (Zhong et al. 2022; Hong et al. 2025). However, the development of depth estimation techniques is hampered by the prohibitive cost of acquiring ground-truth depth annotations. To address this limitation, researchers have explored self-supervised approaches that recover depth cues from video sequences (Godard, Mac Aodha, and Brostow 2017; Godard et al. 2019; Zhou et al. 2017) or stereo image pairs (Godard, Mac Aodha, and Brostow 2017; Wang, Yu, and Gao 2023) using pose or photometric information (the consistency of pixel appearance under different viewpoints).

Conventional self-supervised methods eliminate the need for annotations, but exhibit unreliable performance in adverse weather conditions. Weather particles violate photometric consistency assumptions, impeding real-world deployment. While some approaches investigate weather-invariant feature extraction (Vankadari et al. 2020; Liu et al. 2021), they lack generalization in diverse scenes. Subsequent knowledge distillation methods (transferring knowledge from a teacher model to a student model) (Gasperini et al. 2023; Mao, Liu, and Liu 2024; Tosi, Ramirez, and Poggi 2024) mitigate degradation by using pseudo-labels from clear-weather teachers. However, their independent training prevents effective knowledge transfer.

Contrastive learning (learning by enforcing depth consistency between clean and adverse conditions) presents a viable solution for robust depth estimation. Existing methods enforce depth consistency across scenes (Saunders, Vogiatzis, and Manso 2023; Wang et al. 2024a) but risk collapsing solutions (degenerate output) by directly minimizing depth differences. D4RD (Wang et al. 2024b) mitigates collapse by using diffusion sampling as anchors but requires extra distillation models and computational overhead. WeatherDepth (Wang et al. 2024a) integrates contrastive learning with progressive curriculum learning to balance scenes differences, but depends on preset dataset and complex curricu-

*Corresponding

lum schedule. To overcome these limitations, we propose an efficient plug-and-play contrastive framework that requires neither architectural modifications nor dataset priors.

Inspired by the image restoration method (Wu et al. 2024), we analyze the intrinsic optimization trajectory of self-supervised learning. Intermediate models in different training stages—termed latency models, exhibit progressive convergence toward optimal solutions (Figure 1). This historical state information provides model priors for contrastive learning. By incorporating these self-generated priors, we establish a self-evolution contrastive paradigm that avoids coupling with curriculum learning. Critically, latency models derive from the training process itself and are dataset-agnostic, ensuring generalization across diverse adverse conditions. In addition, severe degradation causes erroneous depth predictions in specific regions, which complicates negative sample identification. To further address weather-induced local depth inconsistencies, we discretize depth maps into intervals and evaluate distributional similarity through probability consistency. This strategy better captures global depth characteristics than pixel-wise regression loss and improves distributional discrimination.

Building on the above analysis, we propose a self-evolution contrastive learning framework for self-supervised robust depth estimation (SEC-Depth). Our approach leverages latency models from historical parameters to generate self-evolution negative samples, while positive and anchor samples derive from the current model under clear and adverse conditions. We implement a dynamic latency model queue updated with recent parameters to maintain state-of-the-art historical representations. Building on our interval-based depth consistency strategy, we propose a self-evolution contrastive loss that constructs anchor-positive-negative triplets from the model's evolving trajectory and enhances robustness by contrasting current outputs with prior suboptimal predictions.

Unlike prior contrastive methods (Wang et al. 2024b,a), our framework does not require baseline modifications and introduces generalization guidance through contrastive learning. The main contributions of this work are:

- We propose a novel self-evolution contrastive framework for robust depth estimation across diverse conditions, compatible with existing self-supervised models without architectural changes.
- We devise a dynamic latency model update strategy and a self-evolution contrastive loss, enhancing representation learning and stable learning in adverse scenes.
- Extensive experiments demonstrate significant improvements across multiple self-supervised tasks and strong zero-shot generalization (evaluation on unseen datasets) on six benchmarks.

Related Work

Self-Supervised Depth Estimation

Self-supervised depth estimation eliminates the reliance on depth annotations by leveraging geometric constraints from video sequences or stereo image pairs. Zhou et al. (Zhou et al. 2017) pioneer this approach using geometric constraints between consecutive frames, establishing the foundation for monocular depth estimation. Subsequent research expands self-supervised learning using stereo pairs (Godard, Mac Aodha, and Brostow 2017; Garg et al. 2016) and videos (Godard et al. 2019; Watson et al. 2021). Recent advances notably improve accuracy through data augmentation (He et al. 2022; Yao et al. 2024), self-distillation (Marsal et al. 2024; Wang, Yu, and Gao 2023), multi-scale feature fusion (Liu et al. 2024b), temporal fusion (Liu et al. 2024a) and stronger network backbones (Zhang et al. 2023; Zhao et al. 2022). Nevertheless, these methods often fail in complex weather conditions where factors such as illumination changes or precipitation violate the *photometric consistency assumption*, limiting practical deployment.

Robust Depth Estimation

Achieving robustness of depth estimation is essential for real-world applications. Early approaches tackle domain adaptation via adversarial learning and image translation (Vankadari et al. 2020; Zhao, Tang, and Sun 2022), yet remain ineffective in adverse conditions (*e.g.*, nighttime, rain). Subsequently, md4all (Gasperini et al. 2023) introduces knowledge distillation for robust training, while the latter works refine it through data augmentation (Tosi, Ramirez, and Poggi 2025; Mao, Liu, and Liu 2024) and improves training strategies (Yan et al. 2025; Jiang et al. 2025). However, student models remain fundamentally constrained by teacher performance. Recent methods employ contrastive learning to enforce consistency between clean and degraded images. Robust-Depth (Saunders, Vogiatzis, and Manso 2023) uses semi-augmented warping and bidirectional contrastive losses. WeatherDepth (Wang et al. 2024a) applies curriculum learning (gradual exposure to harder samples) with progressive adaptation. D4RD (Wang et al. 2024b) integrates multi-level contrastive learning with diffusion models. **Our work** distinguishes itself by proposing a plug-and-play contrastive framework that enhances robustness, which leverages latency models from historical parameters, combined with our interval-based depth consistency strategy.

Method

Preliminaries

Self-supervised depth estimation methods predict a disparity map D by leveraging geometric relationships between a target image I and an auxiliary image I'. Using camera parameters, this disparity D can be converted to a depth map D'. Denoting the depth estimation model as $F : I \to D \in \mathbb{R}^{W \times H}$, the network combines camera intrinsics K and relative pose $T_{I \to I'}$ (obtained from either pose network or extrinsic parameters) to synthesize a warped image $\tilde{I}'$ from I':

$$\tilde{I}' = I' \left\langle \text{proj}\left(D', T_{I \to I'}, K\right) \right\rangle, \tag{1}$$

where $\langle \cdot \rangle$ denotes the pixel sampling operator. The depth prediction is constrained by a photometric reconstruction loss between I and $\tilde{I}'$:

$$L_{ph} = \beta_1 \left(1 - \text{SSIM}(I, \tilde{I}')\right) + \beta_2 \left| I - \tilde{I}' \right|. \tag{2}$$

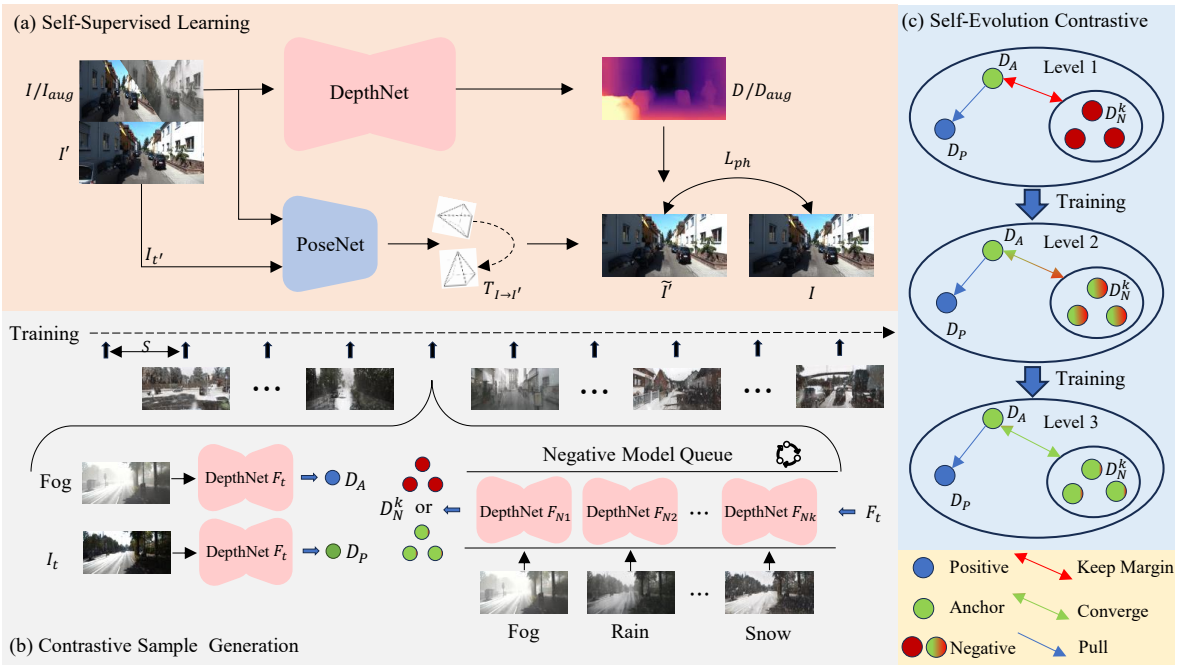

Figure 2: Illustration of our proposed pipeline. (a) Self-supervised learning is conducted on clean images. When augmented samples are introduced, the loss is computed using Equation (4). (b) During training, we maintain a model queue of size j, with parameters updated according to Algorithm 1. (c) As the self-supervised model continues to train, the parameters stored in the model queue gradually converge toward suboptimal states. Our self-evolution contrastive loss is designed to effectively leverage this parametric evolution.

Here, SSIM quantifies structural similarity between images. When processing augmented inputs I_{aug}, the warped image is computed using the corresponding depth D'_{aug} following (Saunders, Vogiatzis, and Manso 2023; Wang et al. 2024a):

$$\tilde{I}' = I' \left\langle \text{proj}\left(D'_{\text{aug}}, T_{I\to I'}, K\right)\right\rangle. \tag{3}$$

Overview

Our self-evolution contrastive learning framework operates independently from the self-supervised depth module. While the base model trains on clean scenes under normal conditions using Equation (1), we introduce a contrastive loss L_c that directly leverages challenging weather conditions to enhance generalization.

Given paired samples (I, I_{aug}) where I is a clean image and I_{aug} is its weather-corrupted counterpart (*e.g.*, rain, snow or fog), both share identical scene content, but differ in appearance. The overall training objective is depicted as

$$L = L_{ph} + wL_c, \tag{4}$$

where w controls the contrastive weight. Crucially, L_{ph} is calculated for both clean (I) and augmented (I_{aug}) images, while L_c is applied *only* when augmented images are processed. As shown in Figure 2, augmented samples I_{aug} are periodically injected into training at fixed intervals S. This allows progressive adaptation to adverse conditions without disrupting the core self-supervised paradigm.

Negative Samples Generation

We maintain a queue of j historical models (called *latency models* or *negative models*) $\{F_{N1}, F_{N2}, \ldots, F_{Nj}\}$, initialized randomly before training. These are updated periodically using an exponential moving average (EMA) of the main model's parameters:

$$\theta_k^* = \omega\theta_k^* + (1-\omega)\theta, \tag{5}$$

where θ_k^* and θ are the parameters of the k-th negative and current main model, updated with a momentum of $\omega = 0.01$.

To ensure that the negative model queue retains up-to-date information from the model during training, we adopt an adaptive update strategy for negative model queue. The detailed update strategy for the negative model can be found in Algorithm 1. The negative model queue is normally updated every T steps. Additionally, when the model queue fails to generate sufficiently diverse negative samples (measured by the variance of their depth differences) for contrastive learning, we will proactively update it. To generate negative samples, we randomly select M (set to j in our work) augmented images I_{aug} and pass each through a randomly assigned negative model F_{Nk}, yielding disparity maps:

$$D_N^k = F_{Nk}(I_{\text{aug}}), \quad k = 1, 2, \ldots, j. \tag{6}$$

These $\{D_N^k\}$ constitute the negative sample set.

Meanwhile, we define the disparity predictions of the self-supervised model on clean and augmented scenes as the anchor and positive sample, respectively:

$$D_A = F_t(I_{aug}), D_P = F_t(I), \tag{7}$$

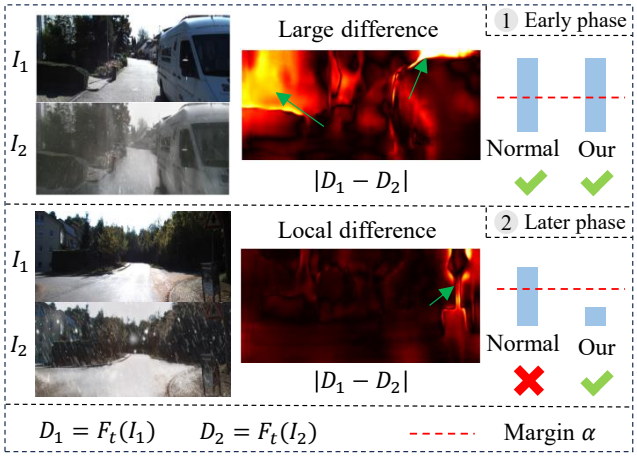

Figure 3: Advantages of our interval-based depth modeling strategy. (1) The model can reliably distinguish samples with significant overall depth differences. (2) Our strategy can better distinguish samples with local depth differences.

where D_A and D_P denote the disparity of the anchor and positive sample. The F_t represents the self-supervised model in training step t, and I_{aug} is the augmented image (*e.g.*, with rain, snow or fog). Since augmented samples are calculated every S steps, the model trains on clean images I during standard self-supervised phases and switches to I_{aug} when contrastive learning is applied. In addition, only D_A (anchor) retains gradients during contrastive loss computation. We denote anchor, positive, and negative samples as A, P, and N respectively in subsequent sections.

Interval-Based Depth Distribution Constraint

Existing contrastive learning methods for robust monocular depth estimation (Saunders, Vogiatzis, and Manso 2023; Wang et al. 2024a) minimize depth discrepancies directly in the depth domain. However, this strategy fails to capture structural relationships when local distortions (*e.g.*, object edges, or weather-induced degradations) dominate global depth distributions. Consequently, models struggle to assess relative depth distributions beyond pixel-wise errors, hindering effective use of negative samples. This limitation is illustrated in Figure 3, where variations in depth distribution impair the model's relational judgment.

To address this issue, we propose a discretized depth modeling strategy that constructs domain-invariant distributional representations. Specifically, we divide the disparity range $[0, 1]$ into N equal bins of width $1/N$, with bin centers $c_n = \frac{n+0.5}{N}$ $(n = 0, 1, \ldots, N-1)$. In this way, the continuous disparity values are converted into a discrete probability distribution. Furthermore, we use a Gaussian kernel $(\sigma = \frac{1}{2N})$ to assign each disparity value d to bins, depicted as

$$w_n(d) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(d-c_n)^2}{2\sigma^2}\right), \tag{8}$$

where $\sigma = \frac{1}{2N}$ denotes the width of the bin. Subsequently, we generate the probability distributions by aggregating weights across all pixels and normalizing to obtain discrete distributions $P_X = [p_X^1, \ldots, p_X^N]$ for anchor (A), positive (P) and negative (N) samples, where $\sum_{n=1}^{N} p_X^n = 1$.

Algorithm 1 Negative Model Queue Update Strategy

Require: Queue Q of size j, current step t, update interval $T_v = 200$, anchor sample a, positive sample p, negative sample set $\mathcal{N}$, current model θ

$n \leftarrow 0$
for each training iteration **do**
  ...
  if $t \bmod T_v = 0$ **then**
    UPDATEQUEUE(Q, θ)
  else if $\mathbb{E}_{n\sim\mathcal{N}}[\mathrm{Var}(D_a - D_n)] < \mathrm{Var}(D_a - D_p)$ **then**
    UPDATEQUEUE(Q, θ)
  end if
end for
function UPDATEQUEUE(Q, θ)
  $Q[n] = \omega Q[n] + (1-\omega)\theta$
  $n \leftarrow (n+1) \bmod M$
end function

Self-Evolution Contrastive Loss

Our loss dynamically adjusts learning objectives using Jensen-Shannon (JS) divergence $(JS(\cdot||\cdot))$ to measure distributional similarity, formulated as:

$$L_c = JS\left(P_A \| P_P\right) + \frac{1}{M}\sum_{M}^{k}\left[\delta\Delta_1^k + JS\left(P_A \| P_N^k\right)\Delta_2^k\right], \tag{9}$$

$$\Delta_i^k = \max\left(\alpha_i - JS\left(P_A \| P_N^k\right), 0\right), i = 1, 2. \tag{10}$$

The dynamically adjusted margin α_1 controls the distance between the anchor and negative samples, allowing the model to adapt its optimization based on their discrepancy. δ is the weight coefficient.

In our experiments, we observe that the self-supervised training process often converges rapidly to a suboptimal state during the early stages. Therefore, we adopt a nonlinear exponential decay strategy to adjust sample distances α_1, allowing the model to better adapt to the evolving learning dynamics throughout training. Then α_1 is:

$$\alpha_1 = ae^{-15\frac{t}{T}} + c, \tag{11}$$

where t is the current training step and T is the total training step, c and a define the range of values of the parameter α_1. We fix the value of α_2 at 0.005 to assess whether the negative sample set reaches a convergent state.

To prevent destabilizing the self-supervised training process—particularly given the model's poor performance in both clear and degraded conditions during the early training stages, we initialize the contrastive loss weight with a small value $(w_s = 0.01)$ and gradually increase it as training progresses. The weight w is defined as:

$$w = \begin{cases} w_s(1 + \max(0, e - e_a)) & e \le e_b \\ w_s(e_b - e_a) & e > e_b, \end{cases} \tag{12}$$

where e is the epoch number, $e_a = 5$ and $e_b = 15$.

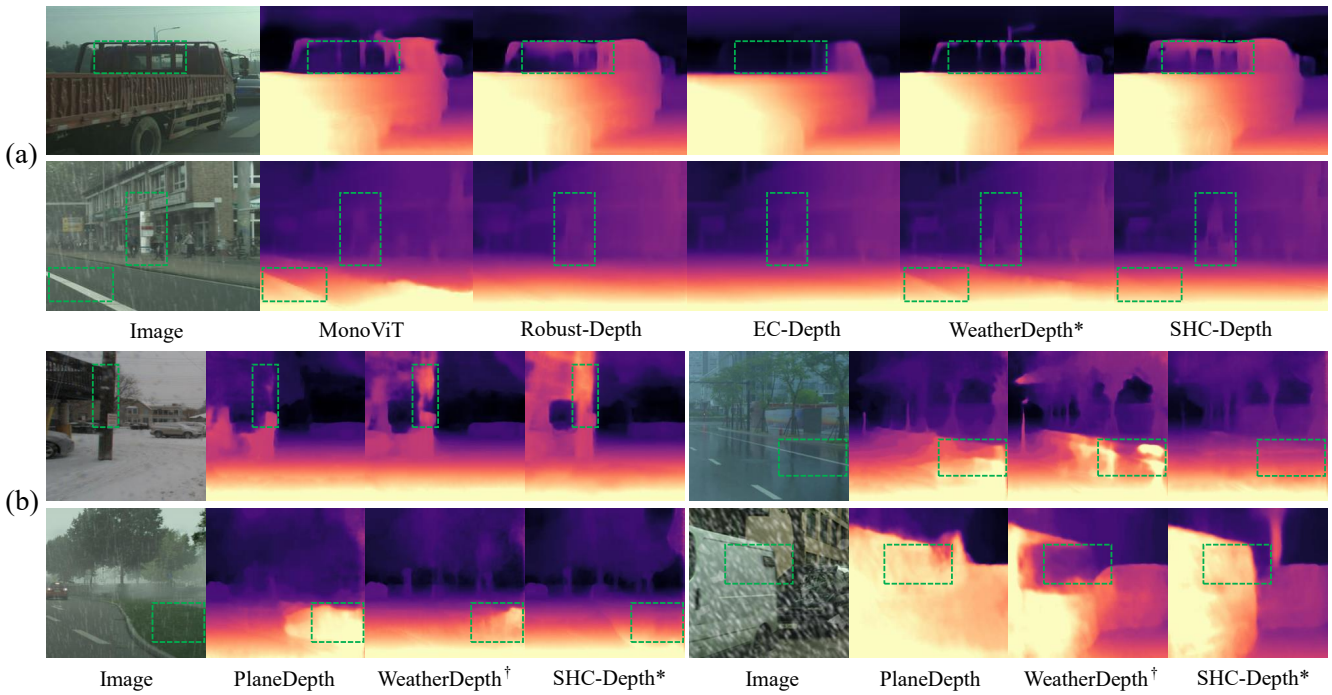

Figure 4: (a) Qualitative results of DrivingStereo and Cityscapes dataset based on the MonoViT baseline. (b) Qualitative results of DrivingStereo and Cityscapes dataset based on the PlaneDepth baseline.

Experiments

Datasets

WeatherKITTI (Wang et al. 2024a) is a synthetic dataset derived from KITTI that includes six weather conditions: two rainy, two foggy, and two snowy scenes. For computing the contrastive loss, we select three representative conditions: mix_rain/50mm, fog/75m, and mix_snow/data. We follow Zhou's split (Zhou et al. 2017), which comprises 39810 training images and 4424 validation images. For testing, we use the Eigen split, consisting of 697 images.

DrivingStereo (Yang et al. 2019) is a real-world dataset. We evaluated the zero-shot performance of the model under challenging conditions using 500 test images each from foggy and rainy scenes.

Cityscapes (Cordts et al. 2016) is a large-scale real-world dataset widely used in autonomous driving. We utilize three publicly synthetic datasets based on Cityscapes: Foggy Cityscapes (Sakaridis, Dai, and Van Gool 2018), SnowCityscapes (Zhang et al. 2021), and RainCityscapes (Hu et al. 2019). Foggy Cityscapes contains 1,525 test images. Compared to the foggy scenes in DrivingStereo and WeatherKITTI, it exhibits more severe visibility degradation. RainCityscapes is derived from the Cityscapes dataset and contains 262 training images and 33 testing images, all of which are normal weather images. Then, three sets of parameters are used to simulate different degrees of rain and fog, attenuation coefficient (0.02,0.01,0.005), (0.01,0.005,0.01) and raindrop radius (0.03,0.015,0.002). In total, 1,188 test images are generated. SnowCityscapes includes three snowfall intensities. From its original 2,000 test images, we select 1,510 samples with available ground-truth depth for zero-shot evaluation.

Dense (Bijelic et al. 2020) is a real-world snowfall dataset. Analysis in (Wang et al. 2024b) demonstrated that the Dense is more suitable for depth estimation tasks than CADC (Pitropov et al. 2021). We followed the processing method described in (Wang et al. 2024b) and obtained 500 test images by sampling 1 in every 3 sequential images.

Implement Details

To validate the effectiveness of our self-evolution contrastive learning framework, we integrate it into two representative self-supervised depth estimation models: MonoViT (Zhao et al. 2022) and PlaneDepth (Wang, Yu, and Gao 2023). Experiments follow the original self-supervised training protocols for clear scenes (KITTI dataset). For the MonoViT baseline, we train for 30 epochs and test at 640×192 resolution. For PlaneDepth baseline, it is trained for 60 epochs using stereo inputs at 1280×384 resolution (first-stage only). Evaluation adheres to established test protocols: WeatherKITTI and DrivingStereo follow (Wang et al. 2024a); Foggy/Snow/Rain Cityscapes follow (Saunders, Vogiatzis, and Manso 2023); Dense follow (Wang et al. 2024b). We set the hyperparameters $\delta = 1e-4$, $a = 0.05$, $c = 0.001$, and the bin count $N = 32$ (balancing runtime and accuracy). For contrastive learning, we use $j = 3$ negative models. Most other hyperparameters (batch size,learning rate and input image resolution, etc.) are the same as their baselines (Zhao et al. 2022; Wang, Yu, and Gao 2023).

Method	AbsRel	SqRel	RMSE	a_1	a_2	a_3
MonoViT	0.120	0.899	5.111	0.857	0.953	0.980
Robust-Depth	0.107	0.791	4.604	0.883	0.963	0.983
EC-Depth	0.113	0.828	4.821	0.871	0.959	0.982
EC-Depth*	0.110	0.790	4.745	0.874	0.960	0.983
WeatherDepth*	**0.103**	**0.738**	**4.414**	**0.892**	**0.965**	**0.984**
SEC-Depth	0.104	0.762	4.473	0.891	0.964	0.983

Table 1: Quantitative results on WeatherKITTI dataset using MonoViT as baseline.

Method	AbsRel	SqRel	RMSE	a_1	a_2	a_3
PlaneDepth	0.158	1.585	6.603	0.753	0.862	0.915
WeatherDepth†	0.099	0.673	4.324	**0.884**	**0.959**	**0.981**
SEC-Depth*	**0.098**	**0.652**	**4.392**	0.883	**0.959**	**0.981**

Table 2: Quantitative results on WeatherKITTI dataset using PlaneDepth as baseline.

Method	AbsRel	SqRel	RMSE	a_1	a_2	a_3
(a) DrivingStereo: Rain						
MonoViT	0.175	2.136	9.618	0.691	0.905	0.973
Robust-Depth	0.166	2.014	9.153	0.755	0.939	0.982
EC-Depth	0.162	**1.723**	**8.478**	0.753	**0.948**	**0.986**
EC-Depth*	0.162	1.746	8.538	0.755	0.947	**0.986**
WeatherDepth*	0.158	1.833	8.837	0.764	0.945	0.985
SEC-Depth	**0.157**	1.820	8.999	**0.766**	**0.948**	0.985
(b) DrivingStereo: Foggy						
MonoViT	0.109	1.204	7.760	0.870	0.967	0.990
Robust-Depth	**0.105**	1.132	7.273	**0.882**	**0.974**	0.992
EC-Depth	0.109	1.107	7.230	**0.882**	**0.974**	0.992
EC-Depth*	**0.105**	**1.061**	**7.121**	0.880	**0.974**	**0.994**
WeatherDepth*	0.110	1.195	7.323	0.878	0.973	0.992
SEC-Depth	**0.105**	1.102	7.346	0.880	0.973	0.992
(c) Foggy Cityscapes						
MonoViT	0.156	1.877	9.598	0.770	0.910	0.967
Robust-Depth	0.127	**1.041**	6.617	0.846	0.966	0.991
EC-Depth	0.145	1.603	8.661	0.792	0.928	0.976
EC-Depth*	0.148	1.600	8.607	0.788	0.929	0.977
WeatherDepth*	0.131	1.214	7.089	0.833	0.959	0.989
SEC-Depth	**0.122**	1.028	**6.262**	**0.860**	**0.972**	**0.992**
(d) Rain Cityscapes						
MonoViT	0.181	0.403	2.158	0.711	0.907	0.959
Robust-Depth	**0.151**	0.236	1.589	**0.788**	0.954	0.987
EC-Depth	0.160	0.340	2.027	0.764	0.930	0.971
EC-Depth*	0.163	0.325	1.982	0.755	0.932	0.973
WeatherDepth*	0.155	0.253	1.708	0.773	0.955	0.985
SEC-Depth	0.154	**0.222**	**1.473**	0.783	**0.967**	**0.989**
(e) Snow Cityscapes						
MonoViT	0.229	2.686	10.512	0.599	0.851	0.942
Robust-Depth	0.158	1.757	8.504	0.766	0.928	0.978
EC-Depth	0.156	**1.650**	8.302	0.773	0.931	**0.980**
EC-Depth*	0.158	1.651	8.275	0.770	0.932	**0.980**
WeatherDepth*	**0.155**	1.672	8.263	**0.779**	**0.933**	0.979
SEC-Depth	0.156	1.712	**8.254**	**0.779**	**0.933**	0.979
(f) Dense Snowy						
MonoViT	0.162	2.063	9.797	0.762	0.904	0.955
Robust-Depth	0.157	1.992	8.945	**0.786**	0.923	0.971
EC-Depth	**0.154**	**1.866**	**8.801**	0.782	0.923	**0.972**
EC-Depth*	0.155	**1.866**	8.828	0.780	0.922	**0.972**
WeatherDepth*	0.157	2.000	9.021	0.781	0.919	0.968
SEC-Depth	0.157	1.991	8.856	**0.786**	**0.924**	0.971

Table 3: Zero-shot evaluation on DrivingStereo, Cityscapes and Dense datasets based on MonoViT baseline.

Comparison Results Based on MonoViT

In this section, we evaluate our method on the WeatherKITTI dataset and perform zero-shot testing on six additional datasets using methods based on the MonoViT baseline. We compare the MonoViT baseline (Zhao et al. 2022), as well as three MonoViT-based robust monocular depth estimation models: WeatherDepth* (Wang et al. 2024a), Robust-Depth (Saunders, Vogiatzis, and Manso 2023) and EC-Depth (Song et al. 2023). EC-Depth* is the second stage model in (Song et al. 2023). We use the pre-training parameters they have already released.

WeatherKITTI Results. We show detailed comparative experiments on the WeatherKITTI datasets in Table 1. Our framework demonstrates a 13.33% decrease in AbsRel errors versus the MonoViT baseline. Compared with the robust alternatives (WeatherDepth* (Wang et al. 2024a), Robust-Depth (Saunders, Vogiatzis, and Manso 2023), EC-Depth (Song et al. 2023)), SEC-Depth achieved competitive results, which indicates the effectiveness of our method.

Zero-shot Results. To further demonstrate the robustness of our model, we perform a zero-shot evaluation on six out-of-distributed datasets. As shown in Table 3, our model achieves significant improvements over the baseline MonoViT under three synthetic and three real-world adverse weather conditions. Compared with existing robust depth estimation methods based on MonoViT, our approach achieves state-of-the-art performance on most datasets. Crucially, it surpasses WeatherDepth* (a contrastive learning benchmark) on the majority of evaluation metrics in 5/6 datasets, underscoring the superiority of our strategy. Visualizations in Figure 4(a) demonstrate detailed depth estimation in rain/fog and robust generalization to heavy rainfall, directly linking our framework's design to improved generalization capability.

Comparison Results Based on PlaneDepth

We evaluate our method in the WeatherKITTI dataset and perform zero-shot testing on six additional datasets using the PlaneDepth baseline (Wang, Yu, and Gao 2023). For distinction, **SEC-Depth* denotes our implementation using the PlaneDepth baseline**. We compare against PlaneDepth and its robust variant WeatherDepth† (Wang et al. 2024a), utilizing their released pre-trained parameters.

WeatherKITTI Results. We show comparisons of our method against the baseline PlaneDepth and the robust WeatherDepth† in Table 2. It is obvious that our SEC-Depth* reduces AbsRel by 37.97% compared to PlaneDepth. It also outperforms WeatherDepth†, indicating advantages when extending our strategy to other depth estimation approaches.

Zero-shot Results. To assess generalization to unseen adverse weather conditions, we evaluate on six real-world and synthetic datasets. As shown in Table 5, our method significantly improves over PlaneDepth across all datasets and exhibits stronger zero-shot capability than WeatherDepth†, confirming its robustness. Figure 4(b) displays the qualitative results, showing that our method accurately predicts the depth of buses in snowy scenes and effectively distinguishes reflections on water surfaces.

Modules				WeatherKIITI						Zero-shot Datasets (Average)					
CL	ID	Δ_1	Δ_2	AbsRel	SqRel	RMSE	a_1	a_2	a_3	AbsRel	SqRel	RMSE	a_1	a_2	a_3
				0.120	0.899	5.111	0.857	0.953	0.980	0.169	1.728	8.241	0.734	0.907	0.964
✓				0.106	0.835	4.544	0.889	0.964	0.983	0.146	1.362	7.207	0.802	0.948	0.983
✓	✓			0.105	0.796	4.508	0.890	0.964	0.983	0.144	1.340	6.913	0.807	0.950	0.984
✓		✓	✓	0.105	0.807	4.523	**0.891**	0.964	0.983	0.144	1.339	6.904	0.808	0.951	0.984
✓	✓	✓		0.104	0.772	4.489	0.890	0.964	0.983	0.143	**1.309**	6.893	0.807	0.951	0.984
✓	✓	✓	✓	**0.104**	**0.762**	**4.473**	**0.891**	**0.964**	**0.983**	**0.142**	1.313	**6.865**	**0.809**	**0.953**	**0.985**

Table 4: Ablations on individual components in SEC-Depth. CL refers to the contrastive learning with degraded sample, ID refers to the Interval-Based Depth distribution constraint and Δ_i refers to our self-evolution contrastive loss.

Method	AbsRel	SqRel	RMSE	a_1	a_2	a_3
(a) DrivingStereo: Rain						
PlaneDepth	0.215	3.659	12.112	0.670	0.889	0.964
WeatherDepth†	0.166	1.874	**8.844**	0.748	0.942	**0.985**
SEC-Depth*	**0.157**	**1.795**	8.976	**0.768**	**0.944**	0.984
(b) DrivingStereo: Foggy						
PlaneDepth	0.122	1.416	8.306	0.847	0.961	0.990
WeatherDepth†	0.123	1.404	7.679	0.859	0.968	0.992
SEC-Depth*	**0.110**	**1.072**	**6.972**	**0.876**	**0.978**	**0.994**
(c) Foggy Cityscapes						
PlaneDepth	0.172	1.922	9.157	0.747	0.918	0.966
WeatherDepth†	0.135	1.014	6.306	**0.873**	0.971	0.992
SEC-Depth*	**0.125**	**0.905**	**5.999**	0.857	**0.974**	**0.993**
(d) Rain Cityscapes						
PlaneDepth	0.233	0.448	2.017	0.607	0.882	0.957
WeatherDepth†	**0.151**	0.195	1.402	0.798	0.967	0.989
SEC-Depth*	0.153	**0.194**	**1.366**	**0.800**	**0.969**	**0.990**
(e) Snow Cityscapes						
PlaneDepth	0.395	5.809	14.803	0.362	0.636	0.805
WeatherDepth†	0.174	1.680	8.250	0.739	0.924	0.972
SEC-Depth*	**0.156**	**1.473**	**7.824**	**0.777**	**0.939**	**0.979**
(f) Dense Snowy						
PlaneDepth	0.175	2.140	9.205	0.754	0.911	0.962
WeatherDepth†	**0.165**	1.877	**8.269**	**0.775**	**0.928**	**0.973**
SEC-Depth*	0.168	**1.861**	8.387	0.768	0.926	0.971

Table 5: Zero-shot evaluation on DrivingStereo, Cityscapes and Dense datasets based on PlaneDepth baseline.

Ablation

We conduct ablation studies on WeatherKITTI and the six zero-shot datasets to validate our framework design. Due to space constraints, we report only MonoViT-based experiments, following the implementation details section.

Ablation on Major Design Components. Table 4 evaluates the contribution of individual components. We augment the baseline with adversarially perturbed samples and contrastive learning (CL), while enforcing direct pixel-level depth alignment between clean and augmented data, which yields significant improvements. Subsequently, we introduce the Interval-Based Depth distribution constraint (ID) strategy (rows 3 and 4), which replaces the pixel-level depth alignment with depth distribution constraints (quantifying depth probability within intervals). Our ID strategy improves the RMSE and δ_1 metrics, demonstrating its superiority over direct alignment in terms of both depth domain alignment and negative sample relation judgment. In addition, by introducing the self-evolution contrastive loss (Δ_1 and Δ_2) to incorporate model priors and strategically select *negative samples*, it further boosts performance on zero-shot datasets. Collectively, these components demonstrate robust generalization in both synthetic and real-world adverse conditions.

Ablation on Negative Step Selection. Table 6 examines the impact of negative step S, which controls the sampling frequency of augmented data in loss computation. Smaller S values increase exposure to challenging augmented samples during training, enhancing robustness but extending training time. In contrast, larger S values reduce augmented sample utilization, consequently degrading generalization to adverse conditions. Taking into account both accuracy and computational efficiency, we select $S = 5$ in experiments.

Method	AbsRel	SqRel	RMSE	a_1	a_2	a_3	Time
(a) WeatherKIITI							
S=1	0.105	0.819	4.521	0.892	0.964	0.983	26.2h
S=5	0.104	0.762	4.473	0.891	0.964	0.983	21.5h
S=10	0.105	0.794	4.525	0.890	0.964	0.983	20.6h
S=20	0.106	0.801	4.544	0.887	0.963	0.983	19.5h
(b) Zero-shot Datasets (Average)							
S=1	0.141	1.329	6.773	0.816	0.954	0.985	26.2h
S=5	0.142	1.313	6.865	0.809	0.953	0.985	21.5h
S=10	0.144	1.336	6.930	0.806	0.951	0.984	20.6h
S=20	0.147	1.392	7.145	0.799	0.947	0.983	19.5h

Table 6: Ablation of negative step selection.

Conclusion

In this paper, we propose SEC-Depth, a novel self-evolution contrastive learning framework for self-supervised depth estimation. Our method implements a dynamic update strategy for historical depth models (latency models capturing prior training states) and constructs triplet samples (anchor, positive, and negative examples) using these models. To resolve ambiguous sample relationships in adverse weather, we transform disparity maps into binned probability distributions (discretized depth intervals for robust distributional comparison). Finally, we design a self-evolution contrastive loss that dynamically adapts optimization targets by contrasting current predictions against divergent outputs from historical models. Extensive validation across multiple datasets and baseline architectures confirms its transferability and effectiveness in zero-shot adverse conditions.

Acknowledgments

This research was financially supported by the National Natural Science Foundation of China (62501189, U23B2009), the Natural Science Foundation of Heilongjiang Province of China for Excellent Youth Project (YQ2024F006) and Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (GML-KF-24-09).

References

Bijelic, M.; Gruber, T.; Mannan, F.; Kraus, F.; Ritter, W.; Dietmayer, K.; and Heide, F. 2020. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *CVPR*, 11682–11692.

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

Garg, R.; Bg, V. K.; Carneiro, G.; and Reid, I. 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 740–756.

Gasperini, S.; Morbitzer, N.; Jung, H.; Navab, N.; and Tombari, F. 2023. Robust monocular depth estimation under challenging conditions. In *ICCV*, 8177–8186.

Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 270–279.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *ICCV*, 3828–3838.

He, M.; Hui, L.; Bian, Y.; Ren, J.; Xie, J.; and Yang, J. 2022. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *ECCV*, 565–581.

Hong, S.; Yue, T.; You, Y.; Lv, Z.; Tang, X.; Hu, J.; and Yin, H. 2025. A Resilience Recovery Method for Complex Traffic Network Security Based on Trend Forecasting. *International Journal of Intelligent Systems*, 2025(1): 3715086.

Hu, X.; Fu, C.-W.; Zhu, L.; and Heng, P.-A. 2019. Depth-attentional features for single-image rain removal. In *CVPR*, 8022–8031.

Jiang, K.; Cao, J.; Yu, Z.; Jiang, J.; and Zhou, J. 2025. Always Clear Depth: Robust Monocular Depth Estimation under Adverse Weather. *arXiv preprint arXiv:2505.12199*.

Liu, J.; Kong, L.; Li, B.; Wang, Z.; Gu, H.; and Chen, J. 2024a. Mono-ViFI: A unified learning framework for self-supervised single and multi-frame monocular depth estimation. In *ECCV*, 90–107.

Liu, L.; Song, X.; Wang, M.; Liu, Y.; and Zhang, L. 2021. Self-supervised monocular depth estimation for all day images using domain separation. In *ICCV*, 12737–12746.

Liu, R.; Zhu, D.; Zhang, G.; Wang, L.; and Li, J. 2024b. Self-supervised Monocular Depth Estimation Based on Hierarchical Feature-Guided Diffusion. *arXiv preprint arXiv:2406.09782*.

Mao, Y.; Liu, J.; and Liu, X. 2024. Stealing stable diffusion prior for robust monocular depth estimation. *arXiv preprint arXiv:2403.05056*.

Marsal, R.; Chabot, F.; Loesch, A.; Grolleau, W.; and Sahbi, H. 2024. MonoProb: self-supervised monocular depth estimation with interpretable uncertainty. In *WACV*, 3637–3646.

Pitropov, M.; Garcia, D. E.; Rebello, J.; Smart, M.; Wang, C.; Czarnecki, K.; and Waslander, S. 2021. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5): 681–690.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126: 973–992.

Saunders, K.; Vogiatzis, G.; and Manso, L. J. 2023. Self-supervised Monocular Depth Estimation: Let's Talk About The Weather. In *ICCV*, 8907–8917.

Song, Z.; Zhu, R.; Wang, C.; Deng, J.; He, J.; and Zhang, T. 2023. EC-Depth: Exploring the consistency of self-supervised monocular depth estimation in challenging scenes. *arXiv preprint arXiv:2310.08044*.

Tosi, F.; Ramirez, P. Z.; and Poggi, M. 2024. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *ECCV*, 236–257.

Tosi, F.; Ramirez, P. Z.; and Poggi, M. 2025. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *ECCV*, 236–257.

Vankadari, M.; Garg, S.; Majumder, A.; Kumar, S.; and Behera, A. 2020. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In *CVPR*, 443–459.

Wang, J.; Lin, C.; Nie, L.; Huang, S.; Zhao, Y.; Pan, X.; and Ai, R. 2024a. Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *ICRA*, 4976–4982.

Wang, J.; Lin, C.; Nie, L.; Liao, K.; Shao, S.; and Zhao, Y. 2024b. Digging into contrastive learning for robust depth estimation with diffusion models. In *ACM MM*, 4129–4137.

Wang, R.; Yu, Z.; and Gao, S. 2023. Planedepth: Self-supervised depth estimation via orthogonal planes. In *CVPR*, 21425–21434.

Watson, J.; Mac Aodha, O.; Prisacariu, V.; Brostow, G.; and Firman, M. 2021. The temporal opportunist: Self-supervised multi-frame monocular depth. In *CVPR*, 1164–1174.

Wu, G.; Jiang, J.; Jiang, K.; and Liu, X. 2024. Learning from history: Task-agnostic model contrastive learning for image restoration. In *AAAI*, 5976–5984.

Yan, W.; Li, M.; Li, H.; Shao, S.; and Tan, R. T. 2025. Synthetic-to-real self-supervised robust depth estimation via learning with motion and structure priors. In *CVPR*, 21880–21890.

Yang, G.; Song, X.; Huang, C.; Deng, Z.; Shi, J.; and Zhou, B. 2019. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *CVPR*, 899–908.

Yao, Y.; Wu, G.; Jiang, K.; Liu, S.; Kuai, J.; Liu, X.; and Jiang, J. 2024. Improving domain generalization in self-supervised monocular depth estimation via stabilized adversarial training. In *ECCV*, 183–201.

Yin, W.; Zhang, J.; Wang, O.; Niklaus, S.; Chen, S.; Liu, Y.; and Shen, C. 2022. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 6480–6494.

Zhang, K.; Li, R.; Yu, Y.; Luo, W.; and Li, C. 2021. Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE Transactions on Image Processing*, 30: 7419–7431.

Zhang, N.; Nex, F.; Vosselman, G.; and Kerle, N. 2023. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *CVPR*, 18537–18546.

Zhao, C.; Tang, Y.; and Sun, Q. 2022. Unsupervised monocular depth estimation in highly complex environments. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5): 1237–1246.

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

Zhong, X.; Tu, S.; Ma, X.; Jiang, K.; Huang, W.; and Wang, Z. 2022. Rainy WCity: A Real Rainfall Dataset with Diverse Conditions for Semantic Driving Scene Understanding. In *IJCAI*, 1743–1749.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 1851–1858.