Pattern Recognition 174 (2026) 112963

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

WPD: Weather prompt driven zero-shot adverse condition depth estimation

Rui Tian [a,1,*], Yongqiang Zhang [b,1], Zhixiang Zhang [a], Man Zhang [a], Zian Zhang [a], Yin Zhang [a], Yongqiang Li [a], Wangmeng Zuo [c]

[a] School of Instrumentation Science and Engineering, Harbin Institute of Technology, Harbin, 150001, China
[b] College of Computer Science, Inner Mongolia University, Hohhot, 010021, China
[c] Faculty of Computing, Harbin Institute of Technology, Harbin, 150001, China

ARTICLE INFO

Keywords:
Adverse condition depth estimation
Zero-shot learning
Style transfer
Weather prompt

ABSTRACT

Existing adverse condition depth estimation methods primarily relied on generative models(*e.g.*, GAN, Diffusion model) to convert the source-domain sunny condition images into those target-domain adverse condition images and keep the invariable depth ground truth at each pixel, resulting in the need of source-target domain image pairs. To tackle this issue, we propose a Weather Prompt Driven zero-shot adverse condition depth estimation method (termed as WPD), which includes Prompt Guided Affine Transformation (PGAT) and Source-Target Visual Consistency (STVC). WPD derives from a CLIP prior that the visual representation is closer to the corresponding text embedding, and an image generation prior that the style of image transfers from source-domain to target-domain can be conducted by using channel-wise mean and variance. Specifically, PGAT is used to estimate unseen target-domain visual representations under the supervision of weather prompt. Meanwhile, STVC is proposed to prevent the estimated unseen target-domain visual representations from deviating too far from source-domain visual representations, thereby preserving the source-domain image content. Extensive experiments demonstrate that the proposed WPD achieves SOTA performance, *e.g.*80.72% on nuScenes-night, 96.09% on nuScenes-rain, 89.40% on RobotCar-night for d_1 metric. Moreover, WPD gains performance improvements 4.90% and 2.02% in d_1 on ORFD-Twili-light and CityScape-foggy compared with the baseline depth estimator. WPD also achieves compelling depth estimation performance under adverse conditions on Driving Stereo-rain and Driving Stereo-foggy for the zero-shot paradigm. Surprisingly, WPD can establish a unified framework across different adverse conditions simply by combining weather prompts. Our code is publicly available on https://github.com/RuiTianHIT/WPD-code.

1. Introduction

The autonomous driving community is increasingly focusing on resolving corner case problems, particularly those related to ensuring driving safety under adverse conditions. Recently, Adverse Condition Depth Estimation (ACDE) [1–4] has emerged as a promising direction. Unlike object detection and semantic segmentation under adverse conditions, acquiring ground-truth depth maps for ACDE relies exclusively on sensor-based measurements(*e.g.*, LiDAR), as manual annotation is infeasible. Under these circumstances, the main challenge in ACDE is that LiDAR no longer provides accurate ground-truth depth maps due to reflections on flooded roads under rainy days or non-textured areas illuminated at night. To mitigate reliance on accurate ground-truth depth maps, unsupervised domain adaption depth estimation methods(*e.g.*, md4all [4] and diffusion fine-tuning protocol [1]) rely on generative models(*e.g.*, GAN or diffusion model) trained on source-target image pairs to convert source-domain images in sunny conditions into target-domain images under adverse conditions as shown in Fig. 1(a). The paradigm ensures that the source-domain images and the translated target-domain images share accurate ground-truth depth maps from the source-domain images, and enhances the robustness of existing depth estimators in recognizing visual patterns under adverse conditions, such as night, rainy and foggy weather. However, these methods depend on sufficient pairs of source-target images to train a generalizable generative model, which increases the burden of target-domain image collection under adverse conditions.

Moreover, the self-supervised depth estimation paradigm [2,3,5,6] is utilized to avoid the inaccurate ground truth of LiDAR during the training phase. This paradigm necessitates well-designed gradient optimization or rigorous geometric constraints; otherwise, the photometric

* Corresponding author.
E-mail addresses: rui.tian.hit@gmail.com (R. Tian), zhangyongqiang@imu.edu.cn (Y. Zhang).
[1] Equal Contribution.

https://doi.org/10.1016/j.patcog.2025.112963
Received 7 August 2025; Received in revised form 24 October 2025; Accepted 20 December 2025
Available online 26 December 2025

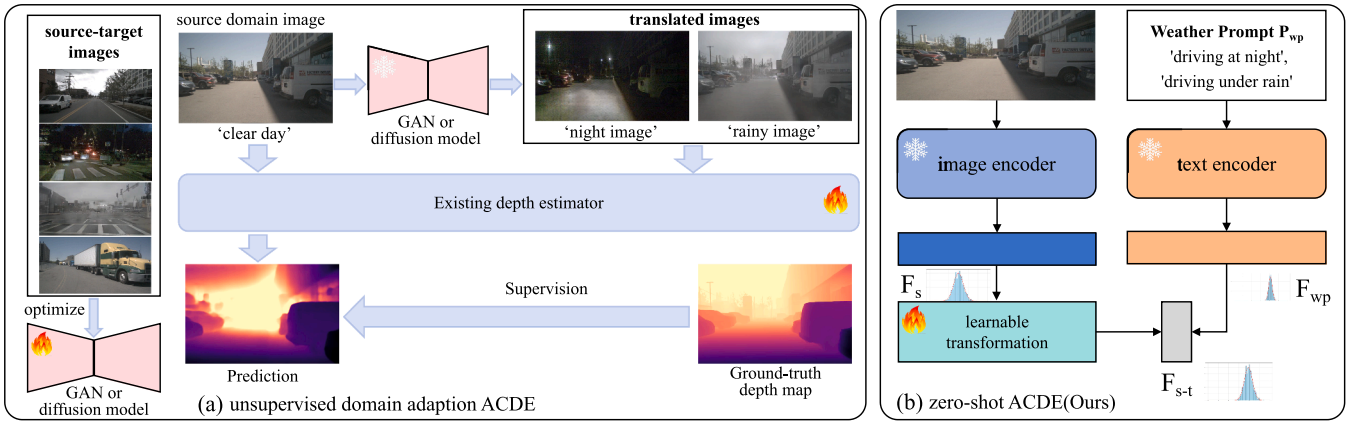

Fig. 1. (a) The unsupervised domain adaption depth estimation methods(*e.g.*, md4all [4] and diffusion fine-tuning protocol [1]) rely on generative models(*e.g.*, GAN or diffusion model) trained on source-target image pairs to convert source-domain images in sunny conditions into target-domain images under adverse weather conditions. (b) The zero-shot adverse condition depth estimation(*i.e.*, our proposed WPD) estimates unseen target-domain visual representations under the supervision of the weather prompt with only accessing source-domain images. The P_{wp}, F_{wp}, F_s and F_{s-t} denote weather prompts, weather prompt embeddings, source-domain visual representations, and estimated unseen target-domain visual representations, respectively. Note that our method operates at the level of visual representations rather than on the entire image.

consistency assumption is easily compromised under complex lighting conditions. Similar to the above generative methods, it is difficult to collect high-quality real-world images from various adverse conditions that vehicles may encounter. Meanwhile, the statistical properties of easily accessible synthetic images often exhibit significant differences from those of real-world images under adverse conditions.

Thus, only given source-domain images under sunny conditions, it is important to develop a robust and generalizable depth estimator that can adapt to various adverse conditions. Our objective is to explore how to estimate pseudo target-domain visual representations with guidance from predefined weather prompts and without target-domain images. Inspired by text-driven domain adaptation [7,8] and text-conditioned feature modulation [9,10], these methods leverage text information as high-dimensional and semantically rich prior information to improve the robustness of the baseline depth estimator in conjunction with the visual modality. This process optimizes internal visual representations and improves depth estimation performance in specific scenarios. Compared to previous text-conditioned feature modulation methods [9,10], which typically rely on cross attention or affine transformation, our approach considers both linear and non-linear transformations to more effectively adapt to adverse conditions, as detailed in Table 7 of Experiment Section. Moreover, existing text-driven domain adaptation technology [7,8] combines visual representations and learnable variables, which is essentially feature augmentation. In contrast, our approach builds on the existing depth estimator and employs the guidance of weather prompts to facilitate feature transfer. This strategy allows the depth estimator to effectively estimate visual representations and adapt to adverse conditions, such as night, rainy, and foggy weather. Additionally, the core of our method essentially is to optimize the learnable depth estimator rather than the learnable variables, which enhances the depth estimator's scalability. As shown in Fig. 6 of Experiment Section, we also compare the depth estimation of the augmentation-based method and our method on nuScenes validation set.

Specifically, in this paper, we present a simple yet effective approach, *i.e.*, weather prompt driven zero-shot adverse condition depth estimation(termed as WPD), as shown in Fig. 1(b). In contrast to unsupervised domain adaptation depth estimation, our method addresses a more challenging zero-shot ACDE task, as it is trained exclusively on source-domain images without accessing any target-domain images. Moreover, our method operates at the level of visual representations rather than on the entire image. The core innovation of WPD lies in two main components: prompt guided affine transformation (PGAT) and source-target visual consistency (STVC). The motivation behind WPD comes from (1) a CLIP prior [8,11,12]: weather prompts can guide the estimation of semantically related visual representations, and (2) image generation prior [13,14]: the style of source-domain images can be adeptly transformed into that of target-domain images using channel-wise mean and variance. The mean and variance of an image effectively capture its statistical characteristics. By adjusting the mean and variance of the source-domain images, the distribution differences between the source-target domain images can be progressively minimized.

Different from the image generation prior, WPD is conducted in the feature space rather than in the image space. Specifically, Prompt Guided Affine Transformation (PGAT) constructs a learnable affine transformation of mean and variance in source-domain images to convert source-domain visual representations into target-domain ones. The core of PGAT is to measure the error for estimated target-domain visual representations and target-domain weather prompts without using target-domain images. We regard the target-domain weather prompt as supervision based on the CLIP prior, and optimize the learnable affine transformation using cosine similarity loss, which minimizes the error of the estimated target-domain visual representations and weather prompt text embeddings. Our designed weather prompts serve as supervisory signals to convert source-domain visual representations into target-domain ones, shifting the paradigm from pixel-aligned supervision(*e.g.*, fully-supervised depth estimator md4all-AD [4] and self-supervised depth estimator Syn2real-depth [6]) to the level of human semantic knowledge. This supervisory signal is more closely aligned with the cognitive essence of target domains(*i.e.*, understanding weather conditions) and perceives visual patterns in the target domains. Our method effectively mitigates the degradation of the supervisory signals caused by adverse conditions and enhances the depth estimator's scalability across multiple adverse conditions.

Moreover, to prevent estimated target-domain visual representations overfitting to weather prompt, Source-Target Visual Consistency (STVC) is designed to bring closer the source-domain visual representations and estimated unseen target-domain ones. This consistency is crucial for preserving the content of source-domain images and preventing overfitting to weather prompts. STVC reinforces consistent representations across modalities, enhancing the generalization ability of WPD to adverse conditions.

As shown in Fig. 2, benefiting from the generalization ability of WPD under various adverse conditions, we observe that high cosine similarity and correlation coefficient of mean and variance are obtained between the estimated unseen target-domain visual representations by WPD and the real target-domain ones under night and rainy conditions. The high

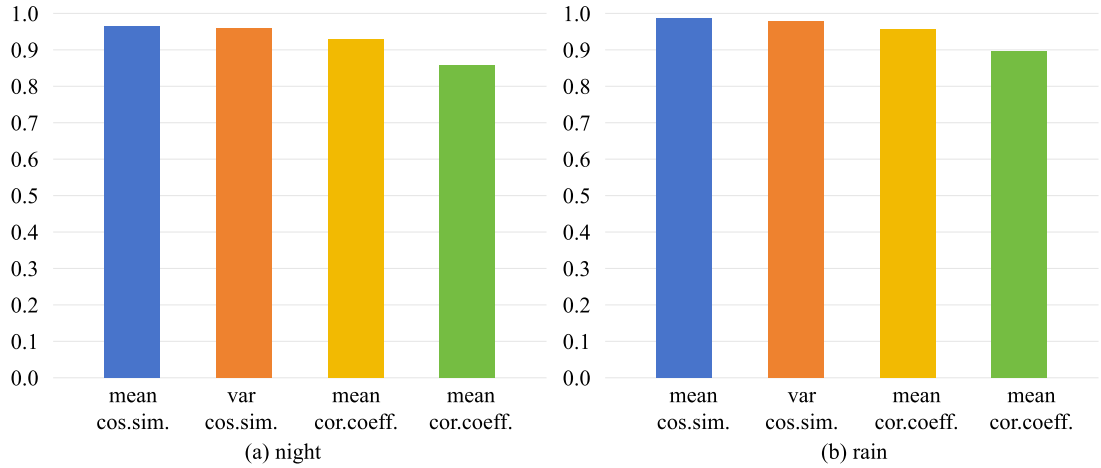

Fig. 2. The comparison for cosine similarity and correlation coefficient of mean and variance between the estimated unseen target-domain visual representations by WPD and real target-domain ones at night and under rain in the first 100 images of nuScenes validation set. From the perspective of contrast consistency, the high similarity and correlation coefficient around 0.9 denote that the distribution ranges of brightness and darkness for the above two features are similar(comparable lighting and shadow effects). From a statistical perspective, the high similarity and correlation coefficient around 0.9 indicate that both the central tendency (*i.e.*, mean) and dispersion (*i.e.*, variance) of the above two features exhibit a highly aligned data distribution pattern.

cosine similarity and correlation coefficient of mean and variance between two visual representations exhibit excellent contrast consistency and the aligned data distribution, which denotes WPD can endow existing depth estimators with the capability to accurately estimate the unseen target-domain visual representations.

In summary, our contributions are summarized as follows:

- We present a novel approach WPD for zero-shot adverse condition depth estimation, which flexibly focuses on a specific weather condition or establishes a unified framework across different adverse conditions by combined weather prompts.
- Prompt Guided Affine Transformation (PGAT) is proposed to optimize the learnable affine transformation for estimating unseen target-domain visual representations without the need for target-domain images. Meanwhile, Source-Target Visual Consistency (STVC) is designed to bring closer the source-domain visual representations and the estimated unseen target-domain ones, preserving the content of the source-domain images.
- Extensive experiments demonstrate the effectiveness of WPD for the adverse condition depth estimation on five popular benchmarks, *i.e.*, nuScenes dataset, Oxford RobotCar dataset, CityScape dataset, ORFD dataset and Driving Stereo dataset. Moreover, WPD has the ability to perceive fine-grained changes(*e.g.*, night and Twililight) in weather through prompt engineering.

The rest of the paper is organized as follows. We first review the related literature in Section 2. In Section 3, the detailed method of our WPD for zero-shot adverse condition depth estimation is presented, which includes the pretraining step, the optimization step with PGAT and STVC, and the fine-tuning for zero-shot ACDE. In Section 4, we first show the main results of WPD and baseline across different adverse conditions for ACDE(*i.e.*, nuScenes-night, nuScenes-day-rain, RobotCar-night, ORFD-Twili-light, CityScape-foggy). Then, we conduct some experiments compared with previous state-of-the-art on the widely used adverse condition depth estimation(*i.e.*, nuScenes-night, nuScenes-rain, RobotCar-night, Driving Stereo-rain and Driving Stereo-foggy), and some ablation study analyzes are performed to validate the effectiveness of each proposed component in our pipeline. Finally, some visualization results are shown to further demonstrate the effectiveness of our WPD for adverse condition depth estimation. This is followed by the conclusions and future work in Section 5.

2. Related work

2.1. Adverse condition depth estimation

Adverse conditions [1–4,15] introduce noise in the pixel correspondences for depth estimation, such as reflections on flooded roads and non-textured areas with vehicle high beams, resulting in erroneous measurements from the LiDAR sensor. A key distinction between depth estimation and downstream perception tasks(*e.g.*, object detection, semantic segmentation) under adverse conditions lies in their dependence on ground-truth supervision. The fully supervised depth estimation fundamentally relies on active sensor like LiDAR for acquiring dense depth annotations, a capability largely beyond human annotation. Additionally, the self-supervised depth estimation adheres to the photometric consistency assumption, and this self-supervised object detection and semantic segmentation are not limited by this assumption. Therefore, existing novel domain adaptation methods based on optimizing multiple tasks [16], frequency information weighting [14] and self-training with pseudo labels [17] are inherently inapplicable to depth estimation under adverse conditions. Existing depth estimation methods [1,4] under adverse condition heavily rely on generative models(*e.g.*, GAN or diffusion model) trained on source-target image pairs to convert source-domain images in sunny conditions into target-domain images under adverse weather conditions. Moreover, some self-supervised depth estimation approaches [2,3,15] avoid inaccurate ground truth with LiDAR to solve the problem of poor illumination or reflections. However, self-supervised depth estimation under adverse conditions violates the assumption of photometric consistency.

Compared with previous work, our method can tackle a specific weather condition and even establish a unified framework for different adverse conditions. During the training phase, WPD only simply introduces the corresponding target-domain text descriptions(*i.e.*, weather prompt) as supervision to estimate unseen target-domain visual representations benefiting from text-driven domain adaptation and text-conditioned feature modulation. Our work is conceptually related to the two technologies described above, but distinct in the internal mechanism. Previous text-conditioned feature modulation methods [9,10] typically rely on cross-attention or affine transformation to perform generative tasks or global image manipulation, while our WPD considers using semantic cues corresponding to weather prompts to control visual feature transfer and explores the ability of linear and non-linear

transformations for ACDE. Moreover, our WPD uses weather prompts to guide feature transfer across domains, whereas existing text-driven domain adaptation techniques [7,8] integrate visual representations with learnable variables as fundamentally a form of feature augmentation.

2.2. Zero-shot depth estimation

Zero-shot depth estimation [4,16,18–20] presents a significant challenge trained on source-domain images and adapted to unseen target-domain images without the need for training or fine-tuning on target-domain dataset. For example, Yang et al. [20] design a foundation depth estimator trained in approximately 62 million images to effectively generalize to an unseen target domain during inference. Hu et al. [19] propose canonical camera space transformation module to solve the metric ambiguity from various camera models and further achieve the zero-shot monocular depth estimation. Piccinelli et al. [16] design a self-promptable camera module predicting the dense camera representation to condition depth features, which achieves the robust zero-shot monocular depth estimation.

However, when high-quality auxiliary datasets are unavailable, the generalization capability of these depth estimators is significantly degraded. To this end, inspired by the CLIP prior [8] that text embeddings can guide the estimation of semantically related visual representations, we propose WPD trained on the sunny condition based on weather prompt guidance to estimate the visual representation of unseen target domain, which effectively adapts to different adverse conditions. To the best of our knowledge, our WPD is the first work to solve zero-shot domain adaptation depth estimation under adverse weather conditions.

2.3. Affine transformation for visual downstream tasks

Affine transformation [11,13,21,22] transforms the source-domain visual features into target-domain ones by calculating the statistics(*i.e.*, mean and variance) of the visual representations in the source-target domain, which has been applied to many downstream visual tasks, such as semantic segmentation [11], depth estimation [21,22], image generation [13,14]. For example, Fahes et al. [11] focus on a specific weather condition to augment low-level source domain features into unseen target domain features based on affine transformations, and achieve robust performance for semantic segmentation. Liu et al. [21] improve the diversity of spatial data augmentation through image affine transformation, and synthesize more virtual camera views by flow-based video frame interpolation. Fu et al. [22] leverage the ability of diffusion priors in generalization and detail preservation to capture geometric details, and combine affine transformation to calculate pseudo metric depth map. Lee et al. [13] design DAFT-GAN, a dual affine transformation generative adversarial network (DAFT-GAN), to minimize information leakage of uncorrupted features for fine-grained image generation.

Previous methods access source-domain and target-domain images, even target-domain ground-truth labels. Compared with previous methods, WPD focuses on solving zero-shot domain adaption task, where the learnable affine transformation is optimized by cosine similarity loss under weather prompt supervision without any target-domain images. We manipulate deep features of the pre-trained text encoder based on CLIP and visual encoder based on the diffusion model. Although affine transformation is not a novel technique, the optimization strategy for affine transformation using prompt weather supervision is novel. Meanwhile, we also explore the comparison of WPD using affine transformation and sophisticated transformation in Section 4.5. The comparison further demonstrates that the linear affine transformation is exceptionally well suited for accurately estimating the visual representations of unseen target-domain images.

3. Method

As previously discussed, WPD aims to solve the problem of zero-shot adverse condition depth estimation by the proposed Prompt Guided Affine Transformation(PGAT) and Source-Target Visual Consistency(STVC). An overview of WPD pipeline is shown in Fig. 3, including a pre-training step, an optimization step, and a fine-tuning step. We select the text-based depth estimator EVP [23] as our baseline. For a better understanding of our WPD, we first describe the EVP training process(*i.e.*, pre-training step of WPD) in Section 3.1. Then, we focus on the method of estimating unseen target-domain visual representations based on PGAT, as well as the strategy to prevent overfitting to weather prompts based on STVC during the optimization step, as described in Section 3.2. Finally, zero-shot adverse condition depth estimation is achieved by fine-tuning the pre-trained EVP on the estimated target-domain visual representations without using any source-domain images in Section 3.3.

3.1. Pre-training step

To enhance the capability of the depth estimator to perceive the surrounding environment, EVP [23] automatically generates image captions P_c leveraging BLIP-2 [24] to capture semantic information about complex road scenarios. Subsequently, visual representations extracted by an image encoder and image caption embeddings obtained by the text encoder integrate into the de-noising UNet with a layer-by-layer cross-attention mechanism to provide explicit guidance for the depth estimation task. The de-noising features of U-Net are fed into Inverse Multi-Attentive Feature Refinement (IMFAR) to capture the enhanced visual representation, and the depth decoder outputs the predicted depth map. The depth estimator is trained with the pixel-wise depth loss based on source-domain sunny condition images as follows:

$$L_{pre} = \sqrt{\frac{1}{T}\sum_i g_i^2 - \frac{\mu}{T^2}(\sum_i g_i)^2} \tag{1}$$

where $g_i = log\ \hat{d}_i - \log d_i$, $\hat{d}_i$ denotes the predicted depth, and d_i denotes the ground-truth depth of source-domain sunny condition images at pixel i. The hyperparameter μ is set to 0.5, consistent with the value used in [23]. T denotes the number of pixels with valid ground-truth values.

3.2. Optimization step with PGAT and STVC

Prompt Guided Affine Transformation(PGAT). The pre-trained depth estimator trained by sunny-domain images only obtains visual representations F_s under sunny conditions, which limits its ability to adapt to unseen target domains. To address this limitation, our objective is to generate robust target-domain visual representations F_{s-t} and effectively adapt to unseen target domains based on weather prompt guidance, without requiring any target-domain images. Specifically, to estimate accurate target-domain visual representations, we take inspiration from CLIP prior and image generation prior. The CLIP prior [8,11,12] indicates that the target-domain visual representation is close to the corresponding text embedding, while the image generation prior [13,14] means that the style of source-domain images can be transformed into that of the target domain images using channel-wise mean and variance. Thus, PGAT constructs an affine transformation of mean and variance for source-domain images to convert source-domain visual representations F_s into target-domain ones F_{s-t} defined in Eq. (2):

$$F_{s-t} = \sigma\frac{F_s - \mu(F_s)}{\sigma(F_s)} + \mu \tag{2}$$

where σ and variance μ are learnable parameters in the affine transformation. To ensure the efficiency of gradient descent, the mean and variance of source-domain visual representations are used to initialize these two learnable parameters. $\sigma()$ and $\mu()$ are two functions that obtain the

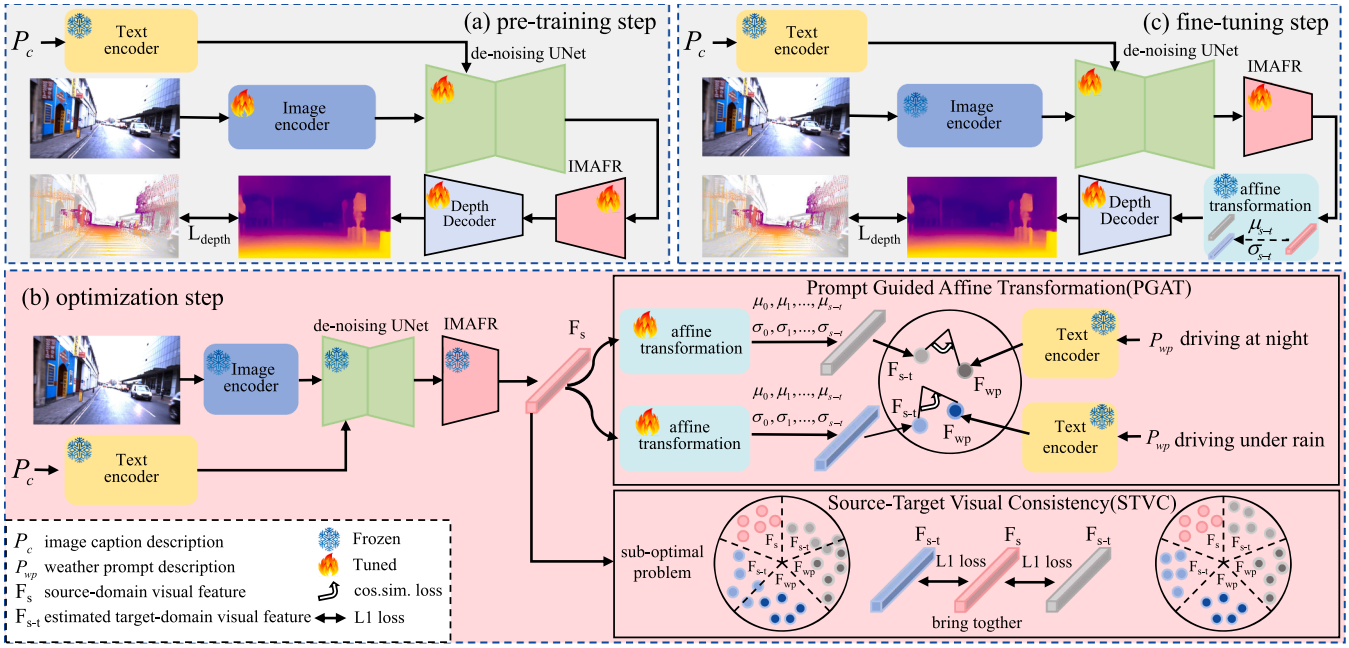

Fig. 3. The overview of the WPD pipeline, where the training process includes the pre-training step, optimization step and fine-tuning step. In inference, the fine-tuned model without the affine transformation is used to achieve zero-shot adverse condition depth estimation.

channel-wise mean and variance of source-domain visual representation F_s. Note that the source-domain visual representation F_s and estimated unseen target-domain ones F_{s-t} have the same size(*i.e.* $F_s, F_{s-t} \in \mathbb{R}^{hwc}$).

The learnable affine transformation based on image generation prior is conducive to solving zero-shot ACDE task since the prior effectively transfers style from source-domain to target-domain and keeps invariable depth ground truth at each pixel with a few parameters. Nevertheless, the target-domain images are not available, resulting in σ and μ lacking the optimization objective. To this end, benefiting from CLIP prior, we consider that the weather prompt corresponding to the target domain can be known in advance. Meanwhile, a few keywords characterizing the weather prompt are provided to optimize the learnable affine transformation. The weather prompts P_{wp} ('driving at night', 'driving under rain', *etc.*) are extracted into text embeddings F_{wp} by CLIP text encoder $T(\cdot)$ as the supervision signal to optimize the affine transformation using cosine similarity loss defined in Eq. (3):

$$L_{PGAT} = 1 - \frac{F_{s-t} \cdot F_{wp}}{||F_{s-t}|| \cdot ||F_{wp}||} \tag{3}$$

To diversify the unseen target-domain text description, we also combine semantic concepts of multiple adverse weather conditions, such as P_{wp} being described as 'driving at night or under rain'. This semantic combination significantly improves the model's ability to generalize across diverse adverse weather scenarios. Considering the limited ability of shallow visual representations to understand global context, we select deep-level and high-dimensional visual representations F_s to implement PGAT as a detailed analysis in Section 4. Note that we also consider the sophisticated transformation to enhance the ability of WPD to estimate the unseen target-domain visual representations as discussed in Table 7 of the Experiment Section.

Source-Target Visual Consistency(STVC). Based on the above process, PGAT enables the pre-trained depth estimator to accurately estimate target-domain visual representation F_{s-t} without requiring target-domain images. However, as the iterative process continues, F_{s-t} easily overfits to the weather prompt F_{wp}. This overfitting can disrupt the depth estimator's ability to generalize to various adverse conditions, resulting in a sub-optimal problem as shown in Fig. 3. To solve this sub-optimal problem, we design Source-Target Visual Consistency (STVC) to bring closer the source-domain and target-domain visual representations, thereby persevering source-domain image contents on estimated unseen target-domain visual representations.

For various choices of consistency losses, we explore L_1 loss and contrastive learning loss to maintain the image content in the source domain during the optimization step. For STVC based on the L_1 loss, we extract the source-domain visual representation F_s for source-domain image I and estimate unseen target-domain visual representations F_{s-t} by PGAT. Utilizing L_1 loss in STVC ensures that estimated target-domain visual representations do not deviate excessively from the source-domain visual representations by Eq. (4):

$$L_{STVC} = ||F_s - F_{s-t}||_1 \tag{4}$$

For STVC based on contrastive learning loss, the target-domain visual representation F_{s-t} is selected as the anchor. The source-domain visual representation F_s is treated as positive samples, while the target-domain weather prompt embedding F_{wp} serves as negative samples. By employing contrastive learning loss, the STVC module effectively brings closer the unseen target-domain visual representations F_{s-t} with the source-domain ones F_s as defined in Eq. (5):

$$L_{STVC} = -log \frac{exp(F_{s-t} \cdot F_s/\tau)}{exp(F_{s-t} \cdot F_s/\tau) + exp(F_{s-t} \cdot F_{wp}/\tau)} \tag{5}$$

where the $\tau = 0.1$ is a temperature parameter. In Section 4, ablation study experiments quantitatively show the effectiveness of STVC based on L_1 loss and contrastive learning loss.

Finally, the overall loss of the optimization step is as follows:

$$L_{opt} = L_{PGAT} + L_{STVC} \tag{6}$$

3.3. Fine-tuning for zero-shot ACDE

During the fine-tuning step, at each training iteration, we estimate target-domain visual representations from source-domain images using trained affine transformation without using any target-domain images. Meanwhile, we freeze the image encoder and the CLIP pre-trained text encoder, then fine-tune the baseline depth estimator using estimated target-domain visual representations and keep invariable depth ground truth at each pixel in the source-domain. Note that we only adjust the style of source-domain visual representation into unseen target-domain

ones, and the pixel-wise depth loss defined in Eq. (1) can still be used to fine-tune the depth estimator. The fine-tuning step introduces the ability to perceive adverse conditions for the pre-trained depth estimator despite the lack of target-domain images. To better understand our training process, we describe the training process in Algorithm 1, which includes the pre-training step, the optimization step, and the fine-tuning step.

In inference, as shown in Fig. 3(c), the fine-tuned model without affine transformation is used to achieve zero-shot adverse condition depth estimation in target-domain images.

Algorithm 1 Pseudo code of WPD.

Require: iteration number N of optimization step,
source-domain visual representation F_s, weather prompt P_{wp}.
//pre-training step
for *epoch* to 10 **do**
pre-train depth estimator EVP by Eq. (1) denoted as M.
end for
//optimization step
for *epoch* to 1 **do**
for $i = 1, 2, \ldots, N$ **do**
calculate channel-wise mean μ^0 and variance σ^0 of F_s.
μ^0, σ^0 initialize learned parameter μ and σ in Eq. (2).
obtain weather prompt text embedding F_{wp} by P_{wp}.
$\mu^i = \mathrm{SGD}(\mu^{i-1}, L_{opt}(\hat{F}_{s-t}, F_{wp}))$, $\sigma^i = \mathrm{SGD}(\sigma^{i-1}, L_{opt}(\hat{F}_{s-t}, F_{wp}))$.
end for
estimate target visual features F_{s-t} by μ^N,σ^N and F_s by Eq. (2)
end for
//fine-tuning step
for *epoch* to 5 **do**
fine-tune M based on F_{s-t} denoted as M^*.
end for
Ensure: M^*

4. Experiments

4.1. Datasets and evaluation metrics

We conduct experiments on five challenging and widely used benchmarks for depth estimation under adverse conditions, *i.e.*nuScenes dataset [25], RobotCar dataset [26], CityScape dataset [27] ORFD dataset [28] and Driving Stereo dataset [29]. For the nuScenes dataset and the Oxford RobotCar dataset, we adopt the split recommended in md4all [4]. **nuScenes dataset** is a challenging dataset with 1000 driving scenes in Boston and Singapore, which contains 15,129 training images and 1570 validation images. Note that in this dataset, training images only include day-clear images without adverse condition images, and validation images are categorized into night and day-rain conditions. **RobotCar dataset** is collected in Oxford and UK by traversing the same route multiple times in a year, which contains 16,563 training images (day-clear images without night conditions) and 708 validation images (only night conditions). **CityScape dataset** covers a variety of urban street scenes in different seasons and sunny weather conditions, including 2975 training images and contains 498 synthetic fog images as testing images. **ORFD dataset** contains woodlands, farmlands, grasslands, and rural areas and captures various weather conditions such as sunny, night, and twilight conditions, where 3960 images under sunny conditions are used as training set and 1666 images under twilight conditions as testing set. **Driving Stereo dataset** includes the real rainy and foggy images, where each condition contains 500 single images for the zero-shot evaluation.

We report on standard metrics and errors up to 50m for RobotCar dataset and 80m for nuScenes dataset, CityScape dataset, ORFD dataset and Driving Stereo dataset, where these standard metrics are defined as follows:

- Abs relative error(absREL): $\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-\hat{y}_i|}{y_i}$;
- Root mean squared error(RMSE): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$;
- Squared Relative Difference(sqlREL): $\frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{y}_i-y_i)^2}{y_i}$;
- Accuracy with threshold t: percentage(%) of $\hat{y}_i$, subject to max $(\frac{\hat{y}_i}{y_i}, \frac{y_i}{\hat{y}_i}) = \mathrm{d} < t$ ($t = 1.25$);

where y_i and $\hat{y}_i$ are the ground-truth depth and estimated depth at pixel i respectively. N is the total number of pixels in the test images.

4.2. Implementation details

In our WPD framework, we select EVP [23] as the baseline depth estimator. All experimental setups are consistent with the baseline unless specifically emphasized. In the pre-trained step, we train EVP [23] for 10 epochs. We use Cosine learning rate adjust strategy with an initial learning rate of 3e-4 and a minimum learning rate of 1e-7. We use the AdamW optimizer with a weight decay of 0.1. The batch size per GPU is set to 3 as consistent with the baseline [23]. During the optimization step, STVC with L_1 loss is used to prevent overfitting to weather prompts. We use IMAFR in the baseline to output visual representations of source-domain images $f_s \in \mathbb{R}^{1024\times12\times12}$. The mean μ and variance σ in the affine transformation are 1024D learnable vectors, and the weather prompt embeddings extracted by the CLIP text encoder are 1024D vectors. We adopt ImageNet templates [12] to encode the weather prompts in text embeddings. We set the base learning rate to 1.0 and train the affine transformation for 20 iterations with batch size of 1 in each GPU using the SGD optimizer. Meanwhile, we set the weight decay wd to 1e-4 and momentum to 0.9. In the fine-tuning step, we freeze the image encoder, CLIP pre-trained text encoder, then further fine-tune the baseline depth estimator with 5 epochs. For a fair comparison with baseline [23], we still use the Cosine learning rate adjust strategy with an initial learning rate of 3e-4 and a minimum learning rate of 1e-7. We use the AdamW optimizer with a weight decay of 0.1. The batch size per GPU is set to 3. In all our experiments, we run the publicly available PyTorch deep learning framework on 2 NVIDIA L20 GPUs.

4.3. Main results

To validate the effectiveness of our WPD, we conduct experiments on nuScenes dataset, RobotCar dataset, ORFD dataset, and CityScape dataset. We introduce the depth estimator EVP [23] into our WPD pipeline, training it during the pre-training stage, optimization step, and fine-tuning step.

Although EVP already performs well under adverse condition depth estimation as shown in Table 1, we can observe that WPD gains performance improvements(*e.g.*RMSE, d_1) compared to the baseline depth estimator EVP [23]. For example, WPD achieves 80.72%, 96.09% and 89.40% in d_1 on nuScenes-night, nuScenes-day-rain and RobotCar-night, which is 2.17%, 0.95, and 1.08% higher than the baseline EVP [23], respectively. Moreover, to further verify the effectiveness of WPD in multiple adverse weather conditions, we conduct experiments on CityScape dataset, *i.e.*, WPD is trained on CityScape-sunny images and tested on CityScape-foggy. We can observe that the performance of d_1 increases further from 88.75% to 90.77% compared to EVP [23], and the RMSE improves by a large margin. Finally, WPD remains effective even in the presence of fine-grained variations in weather conditions. Weather prompts P_{wp} ('driving at Twilight') as the supervision guides WPD to estimate the visual representations of the unseen Twilight domain on the ORFD dataset.

Note that our WPD outperforms the baseline EVP [23] by about 5% absolutely in d_1 from 72.77% to 77.67% on ORFD-Twili-light. The main reason for this improvement performance under multiple adverse conditions is that our WPD indeed estimates target-domain visual features guided weather prompts without the need for additional target-domain

Table 1
Evaluation on nuScenes validation set, RobotCar test set, ORFD test set and CityScape test set.

Method	nuScenes-night		nuScenes-day-rain		RobotCar-night		ORFD-Twili-light		CityScape-foggy	
	RMSE(↓)	d_1(↑)	RMSE(↓)	d_1(↑)	RMSE(↓)	d_1(↑)	RMSE(↓)	d_1(↑)	RMSE(↓)	d_1(↑)
EVP [23]	7.472	78.55	3.355	95.14	2.768	88.32	5.576	72.77	5.780	88.75
WPD	**7.251**	**80.72**	**3.031**	**96.09**	**2.616**	**89.40**	**5.042**	**77.67**	**5.102**	**90.77**

Table 2
Evaluation on nuScenes validation set. WPD with EVP-specific and WPD with EVP-unified denote focusing on a specific weather condition and establishing a unified framework across different adverse conditions under the supervision of combined weather prompts(*e.g.*driving at night or under rain). The zero-shot depth estimators ZoeDepth, Metric3D, Unidepth and relative depth estimators MoGe, Depthanything are pre-trained on mix datasets(*i.e.*, a number of real-world and synthetic datasets) as detail described in [16,19,30] and [20,35]. Moreover, 'mix dataset, day clear' denotes that the WPD with baselines are fune-tuned on nuScenes-day based on pre-trained baselines. The best performance is highlighted in bold.

Method	train data	nuScenes-night			nuScenes-day-rain		
		absREL(↓)	RMSE(↓)	d_1(↑)	absREL(↓)	RMSE(↓)	d_1(↑)
ZoeDepth [30]	mix dataset	0.2580	9.863	54.12	0.2170	9.263	65.57
Metric3D [19]	mix dataset	0.2502	11.474	78.99	0.1001	5.970	90.85
Unidepth [16]	mix dataset	0.1759	7.505	78.13	0.1035	6.262	90.33
WPD with Unidepth(ours)	mix dataset, day-clear	0.1521	7.451	80.42	0.0890	6.017	91.74
Monodepth2 [31]	day-clear	0.2419	10.922	58.17	0.1572	7.453	79.49
PackNet-SfM [32]	day-clear	0.2617	11.063	56.64	0.1645	8.288	77.07
R4Dyn w/o r in [33]	day-clear	0.2731	12.430	52.85	0.1465	7.533	80.59
R4Dyn(radar) [33]	day-clear	0.2194	10.542	62.28	0.1337	7.131	83.91
RNW [34]	day-clear and night	0.3333	10.098	43.72	0.2952	9.341	57.21
md4all-DD [4]	day-clear, Translated night and rain	0.1921	8.507	71.07	0.1414	7.228	80.98
Syn2real-depth[6]	day-clear, Translated night and rain	0.1710	7.689	73.16	0.1270	6.654	82.86
MoGev2 [35]	mix dataset	0.2483	7.570	58.68	0.1649	7.421	79.74
Depthanything [20]	mix dataset	0.2910	11.804	67.10	0.1670	7.867	75.17
WPD with Depthanything(ours)	mix dataset, day-clear	0.2835	10.056	69.03	0.1592	7.717	76.84
SPIdepth [5]	day-clear	0.2100	9.058	67.77	0.1390	7.385	79.25
WPD with SPIdepth(ours)	day-clear	0.1896	8.366	71.29	0.1352	7.128	81.78
DPT [18]	day-clear, night and rain	0.3540	12.875	60.97	0.2370	8.780	66.96
AdaBins [36]	day-clear, night and rain	0.2296	7.344	63.95	0.1726	6.267	76.01
[1] w.DPT[18]	day-clear, Translated night and rain	0.2240	8.375	67.87	0.1990	8.079	72.85
[1] w.md4all-DD[4]	day-clear, Translated night and rain	0.1910	8.433	71.14	0.1470	7.345	79.59
md4all-AD[4]	day-clear, Translated night and rain	0.1821	**6.372**	75.33	0.1562	5.903	82.82
GeoWizard [22]	day-clear	0.1805	7.833	65.94	0.1602	5.260	74.11
marigold [37]	day-clear	0.1833	7.890	65.51	0.1658	5.305	73.67
WPD with marigold(ours)	day-clear	0.1739	7.776	69.03	0.1556	5.206	77.44
EVP [23]	day-clear	0.1612	7.472	78.55	**0.0727**	3.355	95.14
WPD with EVP-specific(ours)	day-clear	**0.1500**	7.251	**80.72**	0.0805	**3.031**	**96.09**
WPD with EVP-unified(ours)	day-clear	0.1519	7.179	80.30	0.0798	3.298	95.63

images, and WPD has the ability to perceive fine-grained changes in weather through prompt engineering(*i.e.*, 'driving at night' or 'driving at Twililight'). Based on the experiments described above, we demonstrate the effectiveness of our proposed WPD trained on only source-domain images(sunny condition) under zero-shot adverse condition depth estimation.

4.4. State-of-the-art comparison

We compare our WPD with other monocular depth estimation methods under adverse conditions, and the comparison results are shown in Table 2. We split Table 2 into three parts according to zero-shot metric depth estimation, relative depth estimation and metric depth estimation containing diffusion-based methods. For each type of depth estimation method, we select at least one method as our baseline and plug it into WPD to further verify the performance of depth estimation under adverse conditions.

In terms of the zero-shot metric depth estimation methods, we can observe that UniDepth pre-trained on extensive real-world and synthetic datasets covering diverse weather conditions still benefits from WPD(78.13% *vs.*80.42% at night and 90.33% *vs.*91.74% in rain). This suggests that WPD effectively transfers visual representations specific to unseen target domains, enabling more effective learning of weather-related representations. For relative depth estimation methods, we select the self-supervised depth estimators Depthanything [20] and SPIdepth [5] as the baselines to integrate into our WPD, denoted as WPD with Depthanything and WPD with SPIdepth. WPD with Depthanything and WPD with SPIdepth improve depth estimation performance(*e.g.*, from 67.10% to 69.03% at night and from 79.25% to 81.78% in rain) under adverse conditions. Meanwhile, WPD with SPIdepth achieves a comparable performance to SOTA self-supervised depth estimators Syn2Real-Depth and R4Dyn with Radar(*e.g.*, 71.29% *vs.*73.16% at night and 81.78% *vs.*83.91% in rain). In terms of metric depth estimation methods, we integrated the diffusion-based depth estimators marigold [37] and EVP [23] into our WPD. WPD with EVP achieves the best results for the majority metrics, where WPD with EVP-specific is trained by one weather prompt and WPD with EVP-unified is trained by multiple weather prompts. For example, our WPD with EVP-specific exceeds md4all-AD [4] by 5.39% (from 75.33% to 80.72%) at night condition and 13.27%(from 82.82% to 96.09%) under rain condition in d_1 metric. Note that WPD with EVP-specific only relies on source-domain images and the target-domain weather prompt, such as 'driving at night' or 'driving under rain'.

To enhance the generalization ability of WPD under various adverse conditions, we establish a unified framework across different adverse conditions (*i.e.*, WPD with EVP-unified) by using 'driving at night or

Table 3
Evaluation on the RobotCar test set, where train data(tr.data) are descried in Table 2. tr.data: Training data, d: day-clear, T:translated in, n: night, a: all. The best performance is highlighted in bold.

Method	tr.data	RobotCar-night			
		absREL(↓)	sqREl(↓)	RMSE(↓)	d_1(↑)
Monodepth2 [31]	d	0.3029	1.724	5.038	45.88
DeFeatNet [38]	a:dn	0.3340	4.589	8.606	58.60
ADIDS [15]	a:dn	0.2870	2.569	7.985	49.00
RNW [34]	a:dn	0.1850	1.710	6.549	73.30
WSGD [3]	a:dn	0.1740	1.637	6.302	75.40
md4all-DD [4]	dT(n)	0.1219	0.784	3.604	84.86
[1] with DPT [18]	dT(n)	0.1330	0.824	3.712	83.95
[1] w. DepthAnything [20]	dT(n)	0.1290	0.751	3.661	83.68
Syn2real-depth [6]	dT(n)	0.1103	0.693	3.455	86.15
EVP [23]	d	**0.0911**	0.416	2.768	88.32
WPD with EVP(ours)	d	0.0999	**0.392**	**2.616**	**89.40**

Table 4
The zero-shot evaluation on Driving Stereo-rain and Driving Stereo-foggy cross weather conditions and cross datasets.

Method	Driving Stereo-rain		Driving Stereo-foggy	
	RMSE(↓)	d1(↑)	RMSE(↓)	d1(↑)
Robust-depth [2]	11.657	59.58	6.876	84.01
Manydepth [39]	8.515	69.24	6.740	83.50
VPD [40]	8.602	68.19	6.694	83.95
EVP [23]	8.397	71.39	6.451	86.03
md4all [4]	8.465	70.35	**6.383**	86.16
Syn2real-depth [6]	**8.004**	72.98	6.422	84.52
WPD(Ours)	8.194	**73.16**	6.390	**86.27**

under rain' prompts, as shown in the last row of Table 2. We can observe that WPD-unified still outperforms the SOTA method (md4all) by a large margin, *e.g.*, 75.33% vs. 80.30% and 82.82% vs. 95.63% at night and under rain for the d_1 metric. The reason for this improvement is that WPD can estimate accurate target-domain visual representations to improve the ability of depth estimation under the adverse condition.

The above experimental results reveal that WPD across various baselines consistently achieves superior depth estimation performance under adverse conditions. Notably, the original EVP inherently incorporates the CLIP text encoder, eliminating the need for additional introduced CLIP text encoder to compute weather prompt embeddings. In contrast, other baselines combined with WPD incur additional computational overhead.

Besides, to demonstrate the robustness of WPD, we also compare WPD with state-of-the-art methods on RobotCar dataset, and the results are shown in Table 3. Similar to above experimental results, we outperform all previous methods. Compared with previous SOTA Syn2real-depth [6], our method successfully improves d_1 from 86.15% to 89.40% under the night condition, further demonstrating the effectiveness of WPD in zero-shot adverse condition depth estimation. Meanwhile, compared with the diffusion-based baseline EVP, WPD further achieves impressive performance(*i.e.*, 88.32% *vs.*89.40% in d_1 and 2.768m *vs.*2.616m in RMSE) on RobotCar test set.

The above experiments validate our WPD's zero-shot ability from a cross-weather perspective. To further validate the zero-shot ability of our WPD, we conduct a more rigorous experiment setting across both different weather conditions and datasets *i.e.*, WPD is trained on nuScenes dataset and tested on Driving Stereo-rain and Driving Stereo-foggy. As shown in Table 4, our proposed WPD achieves the best performance(*i.e.*, 73.16% and 86.27%) in d_1 metric and comparable results(8.194m *vs.*8.004m and 6.390m *vs.*6.383m) in RMSE metric. This consistently strong performance across weather and datasets further demonstrates the robustness of our proposed WPD under zero-shot adverse condition depth estimation.

Table 5
Ablation study results of Prompt Guided Affine Transformation (PGAT) and Source-Target Visual Consistency (STVC) with L_1 loss and contrastive learning loss on nuScenes validation set, where CL. stands for contrastive learning loss.

PGAT	STVC w. L_1	STVC w. CL	nuScenes-night		nuScenes-day-rain	
			RMSE(↓)	d1(↑)	RMSE(↓)	d1(↑)
×	×	×	7.472	78.55	3.355	95.14
✓	×	×	7.426	80.54	3.357	95.63
✓	✓	×	**7.251**	**80.72**	**3.031**	**96.09**
✓	×	✓	7.358	80.61	3.385	95.43

Fig. 4 shows the qualitative results of our WPD against previous SOTA depth estimation methods on nuScenes validation set and RobotCar test set. We can see that our WPD can identify critical elements in scenes under night and rainy conditions, and adjacent objects or the same object are more distinguishable compared with other depth estimation methods. For example, WPD eliminates the artifacts of the car caused by night-time and correctly estimates the complete and continuous high-voltage line, as shown in the 2nd and 3rd rows of Fig. 4. In the 4th row of Fig. 4, our method obtains the sharp estimation for the standing pillar and tree at night. As illustrated in the bottom part of Fig. 4, we also compare the 3D point clouds produced by the proposed WPD and the previous SOTA methods. Our proposed WPD constructs 3D point clouds with fine-grained elements in the scenario and geometric fidelity of objects. The qualitative comparison also shows that WPD is effective for zero-shot adverse condition depth estimation.

4.5. Ablation studies

Effectiveness of PGAT and STVC. As shown in Table 5, although the baseline (*i.e.*, EVP) already performs well under adverse conditions, we can observe that adding PGAT to the baseline improves the performance of the d_1 metric by 1.99% (from 78.55% to 80.54%) and 0.49% (from 95.14% to 95.63%) at night and in rainy conditions, respectively. The reason for this improvement is that WPD with PGAT actually understands the semantic information of the weather prompt and estimates target-domain visual representations by weather prompt guidance without the need for additional target-domain images.

We validate the effectiveness of STVC based on L_1 loss and contrastive learning loss, respectively. As shown in the 2nd and 3rd rows of Table 5, we can see that using STCV with L_1 loss obtains 80.72% and 96.09% in d_1 at night and in rainy conditions, clearly surpassing the baseline and baseline with PGAT. This proves the sub-optimal problem overfitting to weather prompt has been well alleviated based on the STVC with L_1 loss. Meanwhile, we report the performance of using contrastive learning loss in STVC, as shown in the 4th rows of Table 5. We can observe that the d_1 performance under rainy condition decreases from 95.63% to 95.43% compared with the baseline with PGAT, and the performance of STVC with contrastive learning loss is worse than STVC with L_1 loss. We think the reason is that the contrastive learning paradigm, which aims to bring together anchor/positive pairs and push away anchor/negative pairs, impacts the capacity of WPD to utilize the weather prompt as prior knowledge, thus weakens its ability to adapt to adverse conditions. Based on the above experiments, we utilize the L_1 loss in STVC in this paper.

To further validate the effectiveness and robustness of our proposed PGAT and STVC, we evaluate the performance of WPD components on the RobotCar test set. As shown in Table 6, we can observe that adding PGAT to the baseline improves the performance in the d_1 metric by 0.55% (from 88.32% to 88.87%) at night. By incorporating STVC, the performance of WPD increases further from 88.87% to 89.40%, highlighting the benefit of STVC in preventing overfitting to the weather prompt of the target domain. PGAT and STVC understand the semantic information of the weather prompt and prevent the problem of over-

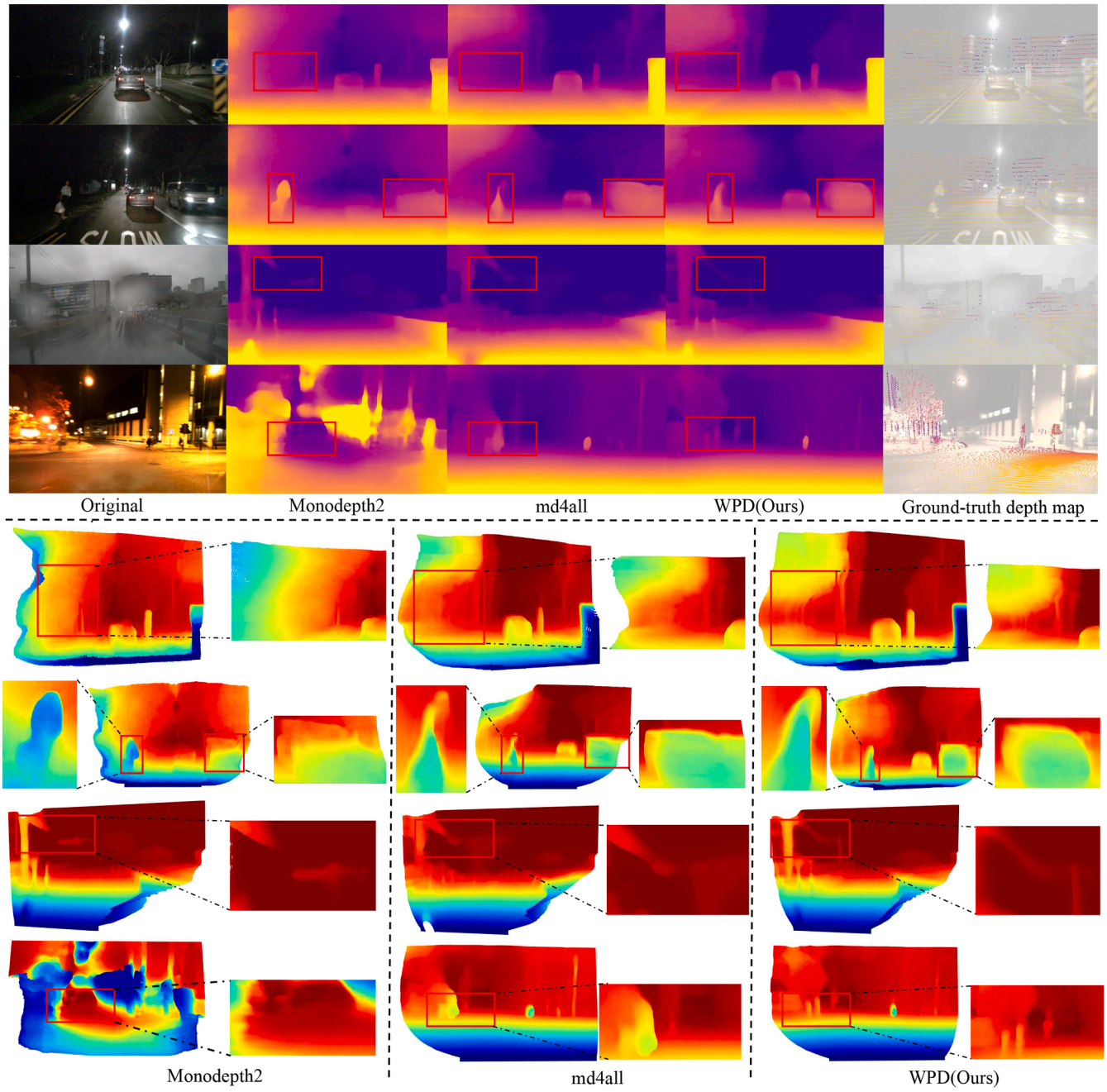

Fig. 4. Qualitative results of WPD against previous SOTA depth estimation methods on nuScenes validation set and RobotCar test set, where the top and bottom parts denote the comparisons of 2D depth map visualization and 3D point cloud visualization with previous SOTA methods. The 3D point cloud constructed by WPD corresponds to the 2D depth map. Best be viewed in color and zoomed in.

fitting to the weather prompt, further validating the effectiveness and robustness of WPD.

Meanwhile, we conduct the ablation visualization of our proposed WPD with PGAT and STCV on nuScenes dataset and RobotCar dataset, as shown in Fig. 5. Our baseline EVP [23] performs well in rainy conditions because rainy weather shares certain characteristics with sunny conditions. However, we found that EVP experiences performance degradation under night conditions, introducing background noise and missed cars. As shown in the third column of Fig. 5, WPD using only PGAT effectively captures the real layout of the cars and eliminates background noise, benefiting from the weather prompt guidance. Moreover, we further bring closer the source-domain visual representations and the estimated unseen target-domain ones by STVC. As shown in the fourth column of Fig. 5, STVC encourages spatial consistency of the visual representation of the original images and the estimated visual representation, preserving the content of original images. Thus, STVC enriches the edge details of the vehicle, person, and traffic sign. Compared with using PGAT alone, WPD using STVC and PGAT gives smooth prediction results and accurate object shapes.

Comparison of sophisticated transformations and affine transformations for WPD. Intuitively, the sophisticated transformation is introduced to enhance the ability of WPD to estimate the unseen target-domain visual representations. To this end, we explore low-rank adaptation matrices(LoRA) as sophisticated transformations and inte-

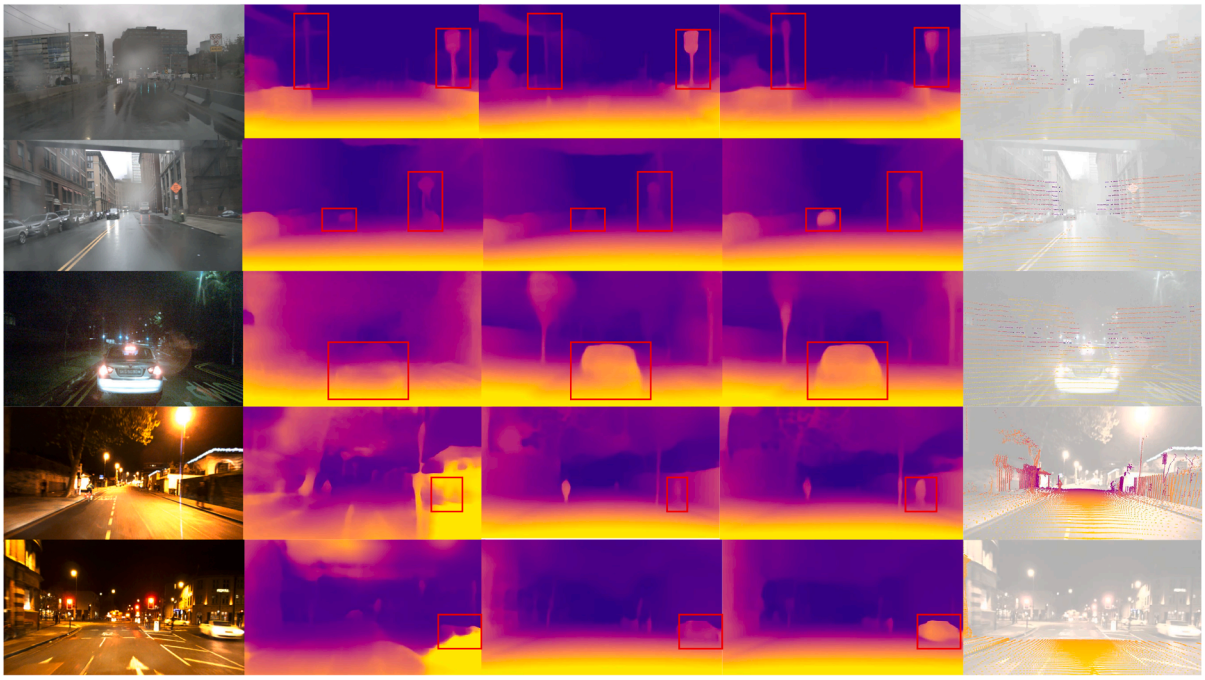

Fig. 5. Ablation visualization of our proposed WPD with PGAT and STCV. The first column denotes the original image. The second, third and fourth columns denote baseline(*i.e.*EVP [23]), WPD only using PGAT, WPD using PGAT and STVC. The final column indicates the depth maps of the ground truth. Best viewed in color and zoomed in.

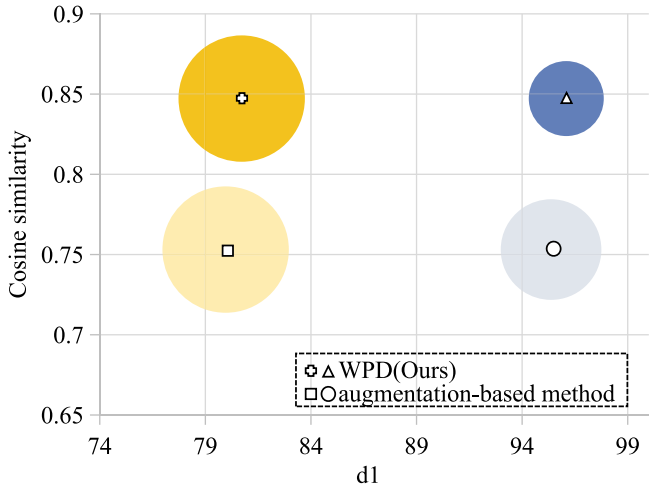

Fig. 6. Comparison of a learnable augmentation based method and our WPD on nuScenes validation set, where the bubble size and x-axis denote RMSE and d_1. The y-axis indicates cosine similarity between the estimated unseen target-domain visual representation and real target-domain ones. Light/dark yellow and light/dark blue indicate the depth estimation at night and under rain, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

grate them into the image encoder to estimate the unseen visual representation of the target domain. As shown in Table 7, we can observe that WPD using affine transformation achieves better performance compared to WPD using LoRA in most metrics. Meanwhile, affine transformation only introduces a minimal increase in parameters for each image(~ 2KB). We believe that affine transformation is sufficient to estimate visual features of unseen target domains in our WPD, achieving excellent performance while introducing a small number of additional parameters.

Comparison of learnable augmentation based method and WPD. To further validate the effectiveness of our WPD, we further compare our method with an augmentation based ACDE method, which benefits from another prior [8] that the source-target difference between language and image should be equal in the embedding space. Following previous work [8], source-domain text descriptions, source-domain visual representations and target-domain text descriptions are used to estimate unseen target-domain visual representations. Compared with this augmentation-based method, our WPD achieves superior depth estimation performance as shown in Fig. 6, *i.e.*the estimated unseen target-domain visual representations by WPD and the real

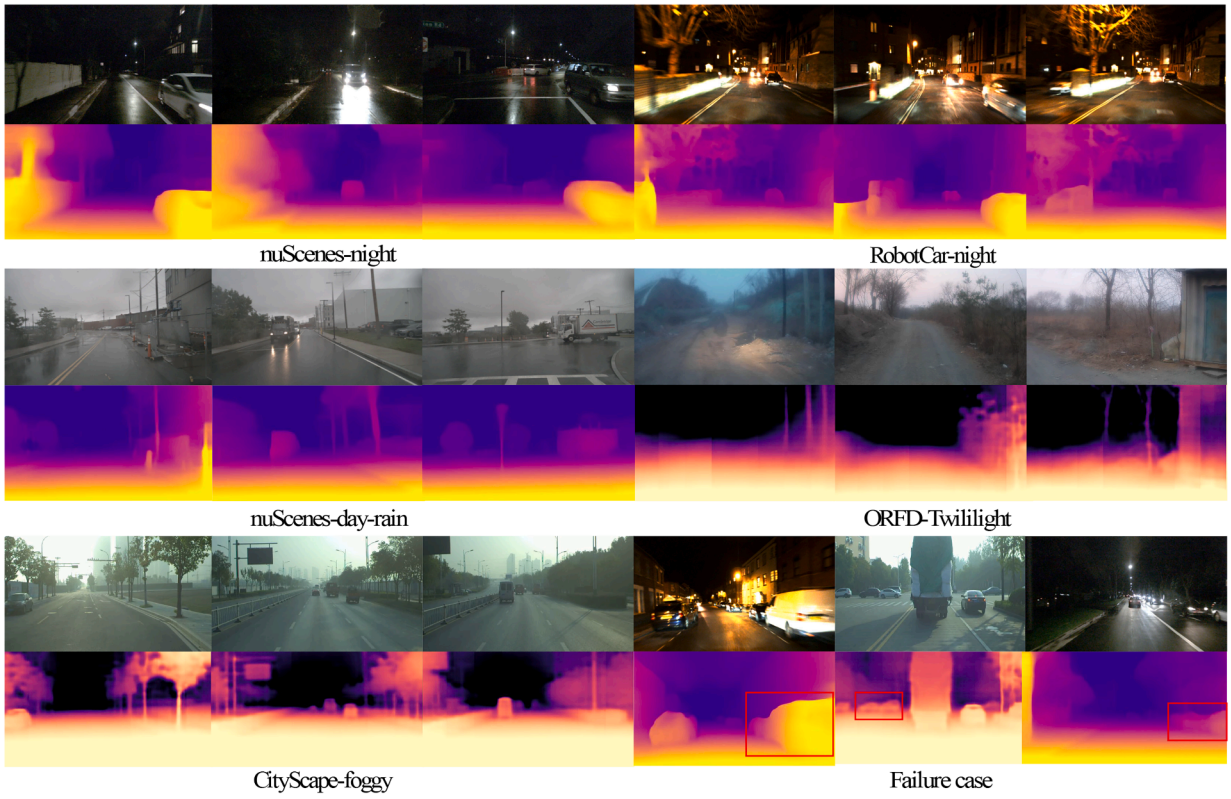

Fig. 7. Qualitative results of our proposed WPD on nuScenes-night, nuScenes-day-rain, RobotCar-night, ORFD-Twililight and CityScape-foggy. Best be viewed in color and zoomed in.

Table 6
Ablation study results of Prompt Guided Affine Transformation (PGAT) and Source-Target Visual Consistenc (STVC) on the RobotCar test set. The performance of the proposed WPD method with or without PGAT and STVC are reported.

PGAT	STVC	night-RobotCar			
		absREL(↓)	sqREL(↓)	RMSE(↓)	d_1(↑)
×	×	**0.0911**	0.416	2.768	88.32
✓	×	0.1076	0.410	2.700	88.87
✓	✓	0.0999	**0.392**	**2.616**	**89.40**

target-domain ones have a higher cosine similarity, and there is a dramatic improvement in the d_1 and RMSE metrics at night and under rainy conditions.

Influence of selecting different layers in PGAT and STVC. We explore the influence of selecting different layers to estimate the target-domain visual representation on RobotCar dataset. As shown in Table 8, we can observe that using the visual representations from IMAFR with 1024D channels obtain unseen target-domain visual representations achieving the best performance 89.40% in d_1 metric. In addition, the visual representation from de-noising UNet with 1320D channels obtains 88.60% in d_1, and 87.70% in d_1 is achieved by using the visual representation from image encoder with 4D channels. Based on above experimental analysis, we conclude that the source-domain visual representations with multiple channels containing richer semantic information, and the position of visual representations close the depth decoder are crucial to estimate unseen target-domain ones in the affine transformation.

Sensitivity analysis of different weather prompt templates P_{wp} and target-domain image captions P_c. We further conduct the sensitivity analysis of different weather prompt templates P_{wp} and image caption P_c for road driving scenarios. Specifically, as shown in the first three rows of Table 9, we further compare the impact of image caption generation(Random *vs.*BLIP-2 [24] *vs.*DAM [41]) on depth estimation performance under adverse conditions. Employing random Gaussian data as textual embeddings and visual representations in cross-attention mechanisms introduces noise into the baseline depth estimator, thus degrading depth estimation performance(*e.g.*, 39.70% *vs.*80.72% in d_1). In addition, our proposed WPD with the image caption p_c from DAM [41] yields enhanced depth estimation performance under adverse conditions. This improvement is attributed to the fact that, compared to BLIP-2 [24], DAM [41] provides WPD with richer and more detailed spatial priors of objects within road driving scenarios, which serve as superior prior information for depth estimation. **For example, DAM: the road is a dark asphalt surface with a glossy finish, reflecting light in certain areas. It has a white dashed line running parallel to the right edge and a solid white line running perpendicular to the dashed line, forming a right angle. The road appears to be wet, with visible reflections and slight variations in texture. BLIP2: a city street at night with a traffic light.**

Meanwhile, we utilize the CLIP score to evaluate the quality of generated captions(*i.e.*, random Gaussian data, DAM-based and BLIP-2-based generated captions) from the perspective of image-text cross-modal matching. We can observe that the DAM-based generated caption achieves the highest CLIP score(*i.e.*, 0.731 *vs.*0.269) compared to the BLIP-2-based generated caption, and the CLIP score of random Gaussian data without any semantic information is almost 0. Moreover, as shown in the last five rows of Table 9, the depth estimation performance exhibits negligible variation across the five prompts, including different wordings, lengths, and languages(English and Chinese). Note that we utilize the Chinese-language CLIP model in place of the vanilla CLIP during the optimization step to better adapt Chinese prompts as shown in the last row of Table 9. Based on the experiment, we further demonstrate

Table 7
The comparison of WPD using affine transformation and sophisticated transformation.

Method	nuScenes-night		nuScenes-rain		RobotCar-night		Parameter(KB)
	RMSE(↓)	d1(↑)	RMSE(↓)	d1(↑)	RMSE(↓)	d1(↑)	
WPD(LoRA)	**7.127**	79.96	3.417	95.37	2.643	89.33	16.384
WPD(affine transfor.)	7.251	**80.72**	**3.031**	**96.09**	**2.616**	**89.40**	**2.048**

Table 8
The influence of selecting different layers to estimate the target-domain visual representation on RobotCar dataset.

image encoder	Denoising UNet	IMAFR	RobotCar-night			
4D channels	1320D channels	1024D channels	absRel(↓)	sqRel(↓)	RMSE(↓)	d1(↑)
×	×	×	**0.0911**	0.4161	2.768	88.32
✓	×	×	0.1105	0.4381	2.838	87.70
×	✓	×	0.1080	0.4177	2.735	88.60
×	×	✓	0.0999	**0.3924**	**2.616**	**89.40**

Table 9
The comparison of different weather prompt templates P_{wp} and image caption P_c for road driving scenarios, where '***' denotes the weather prompts, such as night, rain. The '—' denotes the weather prompts, such as 雨天 (Chinese), 夜晚 (Chinese). The CLIP score of P_c is computed by taking the mean of all images on the nuScenes-night, nuScenes-day-rain and RobotCar-night.

weather prompt templates P_{wp}	image captions P_c	CLIP score of P_c	nuScenes-night		nuScenes-day-rain		RobotCar-night	
			RMSE(↓)	d1(↑)	RMSE(↓)	d1(↑)	RMSE(↓)	d1(↑)
'driving at/under ***'	Random Gaussian data	~ 0.0	10.422	39.70	6.523	61.11	7.187	51.50
'driving at/under ***'	DAM [41]	~ 0.731	**7.007**	**80.96**	**3.009**	**96.25**	**2.583**	**89.64**
'driving at/under ***'	BLIP-2 [24]	~ 0.269	7.251	80.72	3.031	96.09	2.616	89.40
'an image taken on ***' [8]	BLIP-2 [24]	~ 0.269	7.227	79.96	3.117	95.37	2.643	89.33
'an driving view under ***' [42]	BLIP-2 [24]	~ 0.269	7.290	79.85	3.135	94.97	2.699	89.01
'a road of ***' [42]	BLIP-2 [24]	~ 0.269	7.302	79.57	3.149	94.70	2.692	89.02
' 在—行驶 '(Chinese)	BLIP-2 [24]	~ 0.269	7.309	79.70	3.127	95.81	2.710	88.94

Table 10
The comparison of performance improvement from the additional 5 epochs or our WPD.

Method	nuScenes-night			nuScenes-day-rain		
	absREL(↓)	RMSE(↓)	d1(↑)	absREL(↓)	RMSE(↓)	d1(↑)
pre-trained EVP	0.1612	7.472	78.55	**0.0727**	3.355	95.14
ft. only w. 5 epochs	0.1561	7.413	78.95	0.0762	3.530	95.21
ft. w. optim. and 5 epochs	**0.1500**	**7.251**	**80.72**	0.0805	**3.031**	**96.09**

that our proposed WPD is robust to different weather prompt templates and consistently effective regardless of textual differences, and WPD utilizing 'driving at/under ***' yields marginally superior results under adverse conditions.

Does the performance improvement in WPD come from the additional 5 epochs? For the WPD training paradigm, the fine-tuning step incorporates additional 5 epochs. To this end, we explore whether the depth estimation performance improvement under adverse conditions comes from an additional 5 epochs or our WPD. As shown in Table 10, we can see that pre-trained EVP with additional 5 epochs only obtains minimal improvement in d_1 metrics, absREL and RMSE.(*e.g.*from 78.55% to 78.95% in d_1). Instead, following the training paradigm of WPD, the optimization step and fine-tuning step with additional 5 epochs together bring an obvious improvement(*e.g.*from 78.55% to 80.72% in d_1) for pre-trained EVP on the majority of the metrics. Thus, we conclude that the primary enhancements in depth estimation performance are not derived from additional 5 epochs alone, but rather from the optimization step of WPD.

4.6. Qualitative results

Fig. 7 shows the cases of depth estimation results by WPD on nuScenes-night, nuScenes-day-rain, RobotCar-night, ORFD-Twililight and CityScape-foggy. WPD has the ability to correctly perceive weather and capture fine-grained details under multiple adverse conditions, including the thin utility pole, and distant car. Surprisingly, for nuScenes-night, we observe that our model is able to perform well even on night and rain road with reflection and glare. The reflection and glare in the images from the nuScenes-night and nuScenes-day-rain datasets have been successfully removed, providing robust perceptual priors for the control systems of autonomous vehicles. This is attributed to the night and rain prompts (*i.e.*weather prompts) serving as a supervision in WPD replacing the inaccurate depth supervision derived from LiDAR under night and rainy conditions. Meanwhile, WPD preserves the edge details of the tree under Twililight and foggy conditions. We again emphasize that WPD does not access any images from the target domain under adverse conditions. Based on the visualizations, we believe that PGAT indeed estimates the unseen target-domain visual representation, while STVC preserves the content of the original images. Furthermore, WPD still provides reliable depth estimation results even when there are fine-grained changes in illumination (*i.e.*, at night and at twilight).

Finally, we also visualize some failure cases, we find that occluded objects that are close in distance confuse WPD in fine-grained outlines and edges for different objects. Our WPD globally estimates visual representations in these target domains. Therefore, we will subsequently optimize the affine transformation based on patch-level rather than image-level to enhance the generalization ability of WPD.

5. Conclusion

Adverse condition depth estimation frequently struggles with inaccurate LiDAR supervision or violates the photometric consistency assumption. Existing depth estimation methods generally depend on additional target-domain images to capture visual representations of the target domains. In this paper, we propose a WPD framework for zero-shot adverse condition depth estimation based on the weather prompt guidance without using additional target-domain images. Specifically, Prompt Guided Affine Transformation (PGAT) constructs a learnable affine transformation of mean and variance in source-domain images to convert source-domain visual representations into target-domain ones guided by weather prompts using cosine similarity loss. Subsequently, Source-Target Visual Consistency (STVC) brings closer the source-domain visual representations and the estimated unseen target-domain ones to prevent overfitting to weather prompts. Finally, quantitative and qualitative experimental results demonstrate the effectiveness of our WPD and our WPD clearly outperforms state-of-the-art methods on various datasets. Our WPD gains performance improvements across multiple adverse conditions(*e.g.*, night, rain, Twililight and foggy), and has the ability to perceive fine-grained changes(*e.g.*, night and Twililight) under adverse conditions through prompt engineering.

WPD relies on manually crafted prompts to estimate pseudo target-domain visual representations, which enables robust depth estimation under a single weather condition or common mixed weather conditions. However, WPD globally estimates visual representations in these target domains with guidance of manually crafted prompts, which can lead the insufficient fine-grained texture detail at edges of occluded objects. In future work, we will design learnable weather prompts and optimize the affine transformation based on patch-level rather than image-level to enhance the generalization ability of WPD.

CRediT authorship contribution statement

Rui Tian: Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization; **Yongqiang Zhang:** Writing – review & editing, Methodology, Investigation, Funding acquisition; **Zhixiang Zhang:** Validation, Methodology, Data curation; **Man Zhang:** Visualization, Investigation; **Zian Zhang:** Visualization, Software, Methodology; **Yin Zhang:** Visualization, Software; **Yongqiang Li:** Writing – review & editing, Supervision; **Wangmeng Zuo:** Supervision.

Data availability

The data used in this study are publicly available.

Declaration of competing interest

The authors declare they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62206077), the Inner Mongolia Natural Science Foundation for Distinguished Young Scholars (No. 2025JQ009), the Inner Mongolia Talent Development Project for Outstanding Young Talents.

References

[1] F. Tosi, P.Z. Ramirez, M. Poggi, Diffusion models for monocular depth estimation: overcoming challenging conditions, in: European Conference on Computer Vision, Springer, 2024, pp. 236–257.
[2] K. Saunders, G. Vogiatzis, L.J. Manso, Self-supervised monocular depth estimation: Let's talk about the weather, in: IEEE International Conference on Computer Vision, 2023, pp. 8907–8917.
[3] M. Vankadari, S. Golodetz, S. Garg, et al., When the sun goes down: repairing photometric losses for all-day depth estimation, in: Conference on Robot Learning, PMLR, 2023, pp. 1992–2003.
[4] S. Gasperini, N. Morbitzer, H. Jung, N. Navab, F. Tombari, Robust monocular depth estimation under challenging conditions, in: IEEE International Conference on Computer Vision, 2023, pp. 8177–8186.
[5] M. Lavreniuk, A. Lavreniuk, SPIdepth: strengthened pose information for self-supervised monocular depth estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2025, pp. 874–884.
[6] W. Yan, M. Li, H. Li, S. Shao, R.T. Tan, Synthetic-to-real self-supervised robust depth estimation via learning with motion and structure priors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2025, pp. 21880–21890.
[7] Y. Su, Y. Tan, S. An, et al., Semantic-driven dual consistency learning for weakly supervised video anomaly detection, Pattern Recognit. 157 (2025) 110898.
[8] V. Vidit, M. Engilberge, M. Salzmann, Clip the gap: a single domain generalization approach for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 3219–3229.
[9] B. Korbar, Y. Xian, A. Tonioni, et al., Text-conditioned resampler for long form video understanding, in: European Conference on Computer Vision, 2024, pp. 271–288.
[10] M. Pernuš, C. Fookes, V. Štruc, S. Dobrišek, FICE: text-conditioned fashion-image editing with guided gan inversion, Pattern Recognit. 158 (2025) 111022.
[11] M. Fahes, T.-H. Vu, A. Bursuc, et al., Poda: prompt-driven zero-shot domain adaptation, in: IEEE International Conference on Computer Vision, 2023, pp. 18623–18633.
[12] A. Radford, J.W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021, pp. 8748–8763.
[13] K.H.J. Lee, Y. Min, DAFT-GAN: dual affine transformation generative adversarial network for text-guided image inpainting, in: ACM International Conference on Multimedia, 2024, pp. 3275–3283.
[14] H. Gao, B. Ma, Y. Zhang, et al., Frequency domain task-adaptive network for restoring images with combined degradations, Pattern Recognit. 158 (2025) 111057.
[15] L. Liu, X. Song, M. Wang, Y. Liu, L. Zhang, Self-supervised monocular depth estimation for all day images using domain separation, in: IEEE International Conference on Computer Vision, 2021, pp. 12717–12726.
[16] L. Piccinelli, Y.-H. Yang, C. Sakaridis, et al., Unidepth: universal monocular metric depth estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2024, pp. 10106–10116.
[17] Y. Xian, Y.-X. Peng, X. Sun, W.-S. Zheng, Distilling consistent relations for multi-source domain adaptive person re-identification, Pattern Recognit. 157 (2025) 110821.
[18] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer, IEEE Trans. Pattern Anal. Mach. Intell. 44 (3) (2020) 1623–1637.
[19] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, S. Shen, Metric3D v2: a versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 46 (12) 2024 10579–10596.
[20] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, H. Zhao, Depth anything: unleashing the power of large-scale unlabeled data, in: IEEE Conference on Computer Vision and Pattern Recognition, 2024, pp. 10371–10381.
[21] J. Liu, L. Kong, B. Li, Z. Wang, H. Gu, J. Chen, Mono-ViFI: a unified learning framework for self-supervised single and multi-frame monocular depth estimation, in: European Conference on Computer Vision, 2025, pp. 90–107.
[22] X. Fu, W. Yin, M. Hu, K. Wang, Y. Ma, P. Tan, S. Shen, D. Lin, X. Long, Geowizard: unleashing the diffusion priors for 3d geometry estimation from a single image, in: European Conference on Computer Vision, 2024, pp. 241–258.
[23] M. Lavreniuk, S.F. Bhat, M. Müller, P. Wonka, EVP: enhanced visual perception using inverse multi-attentive feature refinement and regularized image-text alignment, 2023. arXiv:2312.08548,
[24] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models, in: International Conference on Machine Learning, 2023, pp. 19730–19742.
[25] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, Nuscenes: a multimodal dataset for autonomous driving, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 11621–11631.
[26] W. Maddern, G. Pascoe, C. Linegar, P. Newman, 1 year, 1000 km: the oxford robotcar dataset, Int. J. Rob. Res. 36 (1) (2017) 3–15.
[27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[28] C. Min, W. Jiang, D. Zhao, J. Xu, L. Xiao, Y. Nie, B. Dai, Orfd: a dataset and benchmark for off-road freespace detection, in: IEEE International Conference on Robotics and Automation, 2022, pp. 2532–2538.
[29] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, B. Zhou, Drivingstereo: a large-scale dataset for stereo matching in autonomous driving scenarios, in: IEEE Conference on Computer Vision and Pattern Recognitionn, 2019, pp. 899–908.
[30] S.F. Bhat, R. Birkl, D. Wofk, P. Wonka, M. Müller, Zoedepth: zero-shot transfer by combining relative and metric depth, 2023. arXiv:2302.12288,
[31] C. Godard, O.M. Aodha, et al., Digging into self-supervised monocular depth estimation, in: IEEE International Conference on Computer Vision, 2019, pp. 3828–3838.

[32] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, A. Gaidon, 3D packing for self-supervised monocular depth estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 2485–2494.
[33] S. Gasperini, P. Koch, V. Dallabetta, et al., R4dyn: exploring radar for self-supervised monocular depth estimation of dynamic scenes, in: IEEE International Conference on 3D Vision, 2021, pp. 751–760.
[34] K. Wang, Z. Zhang, Z. Yan, X. Li, B. Xu, J. Li, J. Yang, Regularizing nighttime weirdness: efficient self-supervised monocular depth estimation in the dark, in: IEEE International Conference on Computer Vision, 2021, pp. 16055–16064.
[35] R. Wang, S. Xu, C. Dai, J. Xiang, Y. Deng, X. Tong, Moge: unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2025, pp. 5261–5271.
[36] S.F. Bhat, I. Alhashim, P. Wonka, Adabins: depth estimation using adaptive bins, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 4009–4018.
[37] B. Ke, A. Obukhov, S. Huang, N. Metzger, R.C. Daudt, K. Schindler, Repurposing diffusion-based image generators for monocular depth estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2024, pp. 9492–9502.
[38] J. Spencer, R. Bowden, S. Hadfield, Defeat-Net: general monocular depth via simultaneous unsupervised representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 14402–14413.
[39] J. Watson, O.M. Aodha, V. Prisacariu, G. Brostow, M. Firman, The temporal opportunist: self-supervised multi-frame monocular depth, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 1164–1174.
[40] W. Zhao, Y. Rao, et al., Unleashing text-to-image diffusion models for visual perception, in: IEEE International Conference on Computer Vision, 2023, pp. 5729–5739.
[41] L. Lian, Y. Ding, Y. Ge, et al., Describe anything: Detailed localized image and video captioning, 2025. arXiv:2504.16072,
[42] F. Zhang, S. You, Y. Li, Y. Fu, Atlantis: enabling underwater depth estimation with stable diffusion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2024, pp. 11852–11861.