

CFDNet: A Generalizable Foggy Stereo Matching Network with Contrastive Feature Distillation

Zihua Liu, Yizhou Li and Masatoshi Okutomi.

Abstract—Stereo matching under foggy scenes remains a challenging task since the scattering effect degrades the visibility and results in less distinctive features for dense correspondence matching. While some previous learning-based methods integrated a physical scattering function for simultaneous stereo-matching and dehazing, simply removing fog might not aid depth estimation because the fog itself can provide crucial depth cues. In this work, we introduce a framework based on contrastive feature distillation (CFD). This strategy combines feature distillation from merged clean-fog features with contrastive learning, ensuring balanced dependence on fog depth hints and clean matching features. This framework helps to enhance model generalization across both clean and foggy environments. Comprehensive experiments on synthetic and real-world datasets affirm the superior strength and adaptability of our method.

I. INTRODUCTION

Stereo matching is a crucial task, where depth is inferred using the disparity between left and right images. Given the affordability of stereo cameras, this technique is popular in fields like Augmented Reality (AR) and autonomous driving. However, in practical situations, especially in the context of autonomous driving, adverse weather conditions such as fog or haze are common. These conditions significantly reduce image clarity, posing challenges for stereo matching. Notably, learning-based stereo-matching networks, typically trained on clean images, exhibit apparent performance degradation in foggy conditions, as illustrated in Fig. 1.

Prior researches [1]–[3] employ networks to mitigate fog’s effects on matching by reconstructing the clean image using the atmospheric scattering model [4]–[6]. These networks simultaneously predict two fog-related hyperparameters: the transmission parameter T and the airlight L_∞ , while estimating disparity [7], [8]. Estimating scattering hyperparameters is challenging with the exponential relationship between transmission and depth. Small errors in transmission estimation can result in significant disparity errors, making training unstable and less generalizable in real-world scenarios.

In contrast to physical prior-based stereo-matching methods for foggy conditions that combine image dehazing, this paper emphasizes enhancing matching feature representations from both clean and foggy images to boost stereo matching across varying conditions. We propose a contrastive feature distillation (CFD) strategy and a network named *CFDNet*. Surprisingly, we discovered that fog does not invariably degrade stereo matching. The fog’s thickness

All authors are with the Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Tokyo, Japan. {zliu, yli, mxo}@ok.sc.e.titech.ac.jp

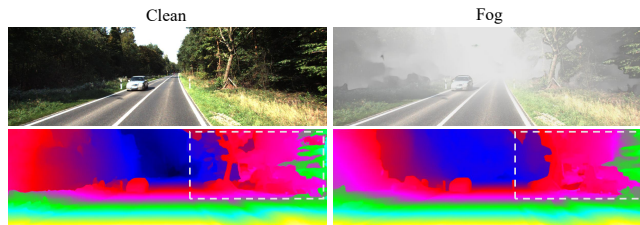


Fig. 1: Clean and foggy left images from stereo image pairs (top), and estimated disparity maps from each pair with state-of-the-art RaftStereo [9] (bottom).

can offer valuable depth cues, where we demonstrate its effectiveness in enhancing matching features significantly with our contrastive feature distillation. Our contributions are summarized as follows:

- **Contrastive Feature Distillation:** We incorporate contrastive learning into feature distillation, ensuring a balanced distilled feature representation between clean and foggy domains. Based on this, we introduce a teacher-student model that optimizes performance across both domains by harmonizing reliance on fog depth cues with clean matching features.
- **Attentive Feature Converter:** We propose an attentive feature converter with stacked pixel and channel-wise attention. It enables fine-grained feature fusion for the *Teacher Model* and feature adaptation for *Student Model* with proposed contrastive feature distillation.
- We conducted comprehensive experiments in both synthetic and real-world settings, utilizing datasets including SceneFlow [10], KITTI [11], and PixelAccurateDepth [12]. Proposed *CFDNet* surpasses leading methods in both foggy and clean environments, underlining its strong efficacy and applicability.

II. RELATED WORK

A. Stereo Matching

Stereo Matching has been an active field and well-studied for decades. Traditional approaches [13]–[17] employ handcrafted patterns for local correspondence matching and exploit spatial context for global optimization. Recently, many learning-based methods [9], [10], [18]–[21] use deep networks to learn discriminative matching features and rely on a cost volume architecture with 2D or 3D convolution aggregation for effective stereo matching. Although these methods have made great progress, since they are often trained on clean images only, they exhibit severe performance

degradation when encountering foggy images, posing challenges for real-world applications like autonomous driving.

B. Stereo Matching in Foggy Conditions.

Enhancing the adaptability of stereo-matching algorithms to various scenarios, especially foggy conditions, has drawn significant research interest. [22] employs a CMOS gated camera alongside a standard stereo camera, aiding depth estimation under severe weather. CATS [3] provides a color and thermal stereo benchmark by applying thermal-thermal matching to improve the robustness of depth estimation in foggy scenes. Caraffa et al. [23] present an MRF model designed for simultaneous stereo matching and dehazing by iterative optimization using α -expansion. Song et al. [24] pioneer a joint-learning framework, combining stereo matching and dehazing using a dual-branch structure with feature fusion. A subsequent version [8] adopts feature attention for a more adaptable feature integration. Shahriar et al. [25] present an integrated solution for 3-D structure that incorporates the visual cues from both disparity and scattering. FoggyStereo [7] delves into depth cues that are hidden in the foggy images, formulating a foggy volume using estimated scattering coefficients to facilitate disparity estimation. While these strategies advance performance under foggy conditions, many necessitate the estimation of scattering coefficients, which can degrade network robustness by introducing uncertainties.

C. Feature Distillation and Contrastive Learning.

Feature distillation [26] commonly serves to transfer knowledge from a teacher model to a student model by mimicking the soft output of the teacher. In this context, Hong et al. [27] introduced a novel method for heterogeneous task imitation. They utilized features derived from clean images as a teaching mechanism, guiding the extraction of features from their hazy counterparts for image dehazing, achieving remarkable results. Moreover, to bridge the domain gap between fog and clean image realms, FIFO [28] implemented a contrastive learning approach. This method compels the network to discern fog-invariant features under the constraints of a contrastive training scheme, thereby exhibiting commendable generalization in segmentation tasks. However, despite the widespread application of both contrastive and distillation learning in representation learning, their potential contributions to stereo matching, particularly under challenging foggy conditions, remain underexplored.

III. PROPOSED METHODS

In this section, we will provide a thorough introduction to our proposed Contrastive Feature Distillation Network (*CFDNet*): the overall architecture of our model is demonstrated in Subsection III-A following the instruction of the attentive feature converter in Subsection III-B. Details about the *Teacher Model* and the *Student Model* can be found in Subsection III-C and Subsection III-D.

A. Network Architecture Overview

The overview architecture of our *CFDNet* with contrastive feature distillation strategy is shown in Fig. 2. The whole network consists of a *Teacher Model* and a *Student Model*. Specifically, the *Teacher Model* takes pairs of clean and foggy images as input, merging features adapted by an attentive feature converter. The *Student Model*, having a similar structure, can accept single domain inputs. It benefits from the guidance of the adaptively fused features by the *Teacher Model*. For robust performance in both hazy and clean domains for the *Student Model*, we introduce a contrastive learning strategy that balances reliance on fog depth hints and matching features, preventing over-dependence on either one. When inference on either clean or foggy images, we utilize our pre-trained *Student Model* for estimating the disparities for real-world applications.

B. Attentive Feature Converter

Motivated by recent advancements in image dehazing [29]–[31], discrepancies between features extracted from foggy and clean images often manifest in both spatial and channel dimensions. Fig. 3 illustrates our attentive feature converter architecture. Similar to the FFA module proposed in [30], our proposed feature converter takes the initial feature F from the feature extractor as input. It employs consecutive pixel-wise attention (PA) and channel-wise attention (CA) modules to synergistically modify features in spatial and channel dimensions. This process can be formulated as:

$$\tilde{F} = F + (F + \text{Conv}(F)) \otimes PA \otimes CA, \quad (1)$$

where the $\tilde{F} \in \mathbb{R}^{H \times W \times C}$ signifies the converted feature, while PA and CA represent pixel-wise and channel-wise attention map, respectively. We utilize a 3×3 convolution layer with sigmoid activation to generate a spatial attention map PA of dimensions $\mathbb{R}^{H \times W \times 1}$. Similar operations are conducted for the $CA \in \mathbb{R}^{1 \times 1 \times C}$, except involving a global average pooling:

$$PA = \sigma(\text{Conv}(F')) \quad (2)$$

$$CA = \sigma(\text{Conv}(\text{avgpool}(F'))) \quad (3)$$

where the F' represents the input feature and σ stands for sigmoid function.

As depicted in Fig. 2, the attentive feature converter assumes distinct roles within the *Teacher* and *Student Models*. In the *Teacher Model*, the attentive feature converter facilitates the adaptively fused of both fog and clean features in spatial and channel domains, thereby enhancing feature representation learning. In the *Student Model*, the attentive feature converter mainly focuses on alleviating the feature gap between the fog feature $F^{S, \text{fog}}$ and clean feature $F^{S, \text{clean}}$ with our contrastive feature distillation.

C. Teacher : Feature Fusion Stereo Matching Network

A simplified RaftStereo [9] is used for disparity regression in our *Teacher Model*. We first extract features $F^{T, \text{clean}}$ and $F^{T, \text{fog}}$ from paired clean and foggy stereo images by

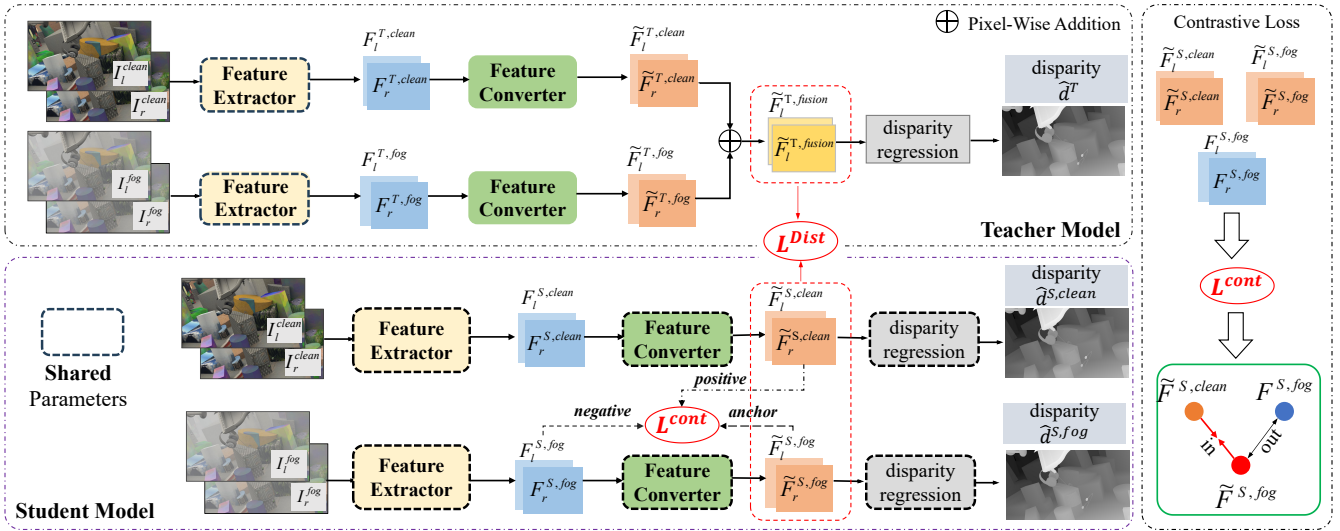


Fig. 2: Overall architecture of our proposed *CFDNet* with contrastive feature distillation.

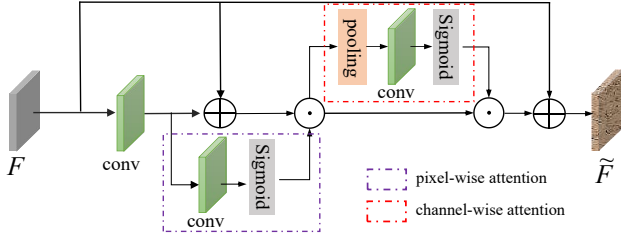


Fig. 3: Attentive Feature Converter. A feature aggregation module with pixel-wise and channel-wise attention layers.

a weight-sharing feature extractor, as depicted in Fig. 2. Notably, though fog affects clarity, it offers depth cues beneficial for disparity regression from a mono-depth aspect. This is rooted in the inverse relationship between fog density and scene depth. Hence, in challenging situations like occlusions, foggy images can provide crucial guidance absent in clean images, while clean images provide vital structural information for matching. As both features have benefits and are complementary to each other, we adaptively fuse them using our attentive feature converter, as depicted in Fig. 2. This fusion is represented as:

$$\tilde{F}^{T, fusion} = FC(\tilde{F}^{T, clean}) + FC(\tilde{F}^{T, fog}), \quad (4)$$

where $\tilde{F}^{T, fusion}$ signifies the adaptively fused feature, and FC denotes the attentive feature converter. We then utilize this feature for disparity estimation. Ablation study result in Table Ion the highlights its superior effectiveness. Consequently, this feature serves as a supervision signal for feature distillation in the *Student Model*.

The *Teacher Model* is trained using the ground truth disparity map, adhering to the sequential loss from [9]:

$$L_{disp}^T = \sum_{i=0}^N \gamma^{N-i} \|d^{gt} - \hat{d}_i^T\|, \quad (5)$$

where i is the current iteration, N represents the total

iteration count (set to 12), and a decay weight γ of 0.95.

D. Student : Parallel Stereo Matching Network with Contrast-Balanced Distillation.

The adaptive fusion of fog and clean features produces a powerful feature, $\tilde{F}^{T, fusion}$, optimal for stereo matching. However, in real-world scenarios, simultaneous acquisition of both clean and hazy image pairs for inference is impractical. Therefore, we introduce a *Student Model* that distills the features from the *Teacher Model* as depicted in Fig. 2. It has a similar architecture to the *Teacher Model* but can accept single domain inputs during inference.

In the training phase, the *Student Model* processes paired clean and hazy images similarly, yielding initial features $F_r^{S, clean}$, $F_r^{S, fog}$ and converted ones $\tilde{F}_r^{S, clean}$, $\tilde{F}_r^{S, fog}$. Contrary to the *Teacher Model*, which combines converted clean and fog features, the *Student Model* enhances feature robustness through contrastive learning incorporated distillation. This approach maintains a balanced emphasis on fog depth hints and matching features during distillation by contrastive learning, avoiding excessive reliance on either.

In the contrastive feature distillation framework outlined in Fig. 2, the features $\tilde{F}^{T, fusion}$ from the *Teacher Model* is used to supervise the converted fog feature $\tilde{F}^{S, fog}$ and clean feature $\tilde{F}^{S, clean}$. Meanwhile, the contrastive learning L^{cont} aims for feature consistency between $\tilde{F}^{S, fog}$ and $\tilde{F}^{S, clean}$ and to encourage difference between $\tilde{F}^{S, fog}$ and $F_r^{S, fog}$.

Intuitively, the distillation loss drives both the converted fog and clean features to closely resemble the adaptively fused features from the *Teacher Model*. However, using the shared feature extractor and converter for both domains can lead to unbalanced performance, potentially making the foggy domain easier to learn and generalize badly on the clean domain. To counteract this, the triplet loss first promotes consistency between the converted clean and foggy features, ensuring balanced proficiency in both domains. Moreover, it differentiates between initial and converted

fog features, preventing the converter from being trivial. preventing the converter from being trivial.

Concretely, for the feature distillation, we minimize the L1 distance between the teacher feature and the student features:

$$L^{Dist} = \left\| \tilde{F}^{T,fusion} - \tilde{F}^{S, clean} \right\| + \left\| \tilde{F}^{T,fusion} - \tilde{F}^{S,fog} \right\|, \quad (6)$$

Regarding contrastive learning, we use the Triplet Loss [32]:

$$L^{Cont} = \max(D(\tilde{F}^{S,fog}, \tilde{F}^{S, clean}) - D(\tilde{F}^{S,fog}, F^{S,fog}) + m, 0), \quad (7)$$

where the m is the distance threshold which in our case equals 1.0. D is the distance function defined as:

$$D(F_1, F_2) = \frac{1}{C} \sum_{i=0}^C (\|F_{1,i}^{norm} - F_{2,i}^{norm}\|_2^2), \quad (8)$$

where C means channel numbers and F^{norm} denotes feature normalized with respect to channel dimension. The *Student Model* is also supervised by the disparity loss defined as:

$$L_{disp}^S = \sum_{i=0}^N \gamma^{N-i} \|d^{gt} - \tilde{d}_i^{T, clean}\| + \sum_{i=0}^N \gamma^{N-i} \|d^{gt} - \tilde{d}_i^{T, fog}\|. \quad (9)$$

The final loss of the *Student Model* is the weighted summation of disparity loss and feature loss:

$$L_{student} = L_{disp}^S + \lambda_1 L^{Dist} + \lambda_2 L^{Cont}, \quad (10)$$

IV. EXPERIMENTS

A. Datasets

We evaluate our methods on multiple public benchmark datasets including SceneFlow [10], KITTI [11], and PixelAccurateDepth [12]. As the training of the proposed network requires pair clean and foggy images, we use the atmosphere scatter equation [5] to render the foggy using the ground-truth disparity for SceneFlow and the KITTI dataset. More details about the foggy scenes synthesis can be seen in Subsection IV-B. The SceneFlow is a synthetic dataset containing 39823 stereo image pairs with random flying objects. The KITTI dataset comprises real-world scenes that have sparse ground-truth disparity captured using LiDAR. The PixelAccurateDepth [12] dataset is a real-world dataset that includes stereo images with 17 levels of artificial fogs and dense ground-truth disparity annotations.

B. Foggy Images Synthesis

For the foggy scenes synthesis, we use the atmosphere scattering function following the previous methods [7], [8], [24] using clean images and its corresponding depth map:

$$I(x) = J(x)T(Z_x) + L_\infty(1 - T(Z_x)), \quad (11)$$

where $I(x)$ is the rendered foggy images, $J(x)$ is the original clean images, and L_∞ stands for the atmospheric light of the given scenes. $T(Z_x)$ is the transmission term indicating the attenuation from an object to the camera which is commonly measured by the Beer-Lambert-Bouguer law [33]:

$$T(Z_x) = e^{-\int_0^{Z_x} \beta(z) dz}, \quad (12)$$

where the Z_x is the given depth at location x , $\beta(\cdot)$ is the attenuation coefficient which reflects the overall concentration of fog. We leverage dense ground truth disparity maps

during the synthesis procedure for the SceneFlow dataset. However, the KITTI 2012 & 2015 datasets comprise solely sparse ground truth annotation. Therefore, we adopt a similar approach to [7] and employ the pre-trained LEAStereo [34] model for generating pseudo-dense disparity maps, which are then employed in foggy data synthesis.

C. Implementation Details

We implemented our *CFDNet* by PyTorch trained with two NVIDIA 3090 GPUs. For the SceneFlow dataset, we trained the Teacher Model for 30 epochs using a batch size of 8 with an initial learning rate of 4e-4 following a step learning rate decay strategy. After that, we train the Student Model with the Teacher Model's intermediate feature for another 30 epochs with distilling loss and triplet loss. The λ_1 and λ_2 in Eq 10 are both set to 0.1 in our cases. As for the KITTI 2015& KITTI 2012 dataset, the ground truth disparity for the benchmark test set is not available for foggy image rendering and testing. Hence, we use the KITTI 2015 training set as training data and fine-tune our SceneFlow pre-trained model with an initial learning rate of 1e-4 for 500 epochs. Then, we use the KITTI 2012 training set as test data to evaluate the performance. For the evaluation of the PixelAccurateDepth dataset, we directly used the model pre-trained on the KITTI 2015 dataset to test the real-fog generalization performances.

D. Evaluation Metrics.

For evaluation on the SceneFlow dataset, we report the standard end-point error (EPE) and the P1-value (outliers percentage) for both clean images and foggy images. As for the KITTI dataset, besides the EPE error, we follow the official evaluation and compute D1 error, which indicates the ratio of outliers with errors bigger than 3 pixels. As for the PixelAccurateDepth dataset, we follow the metrics in [12].

E. Ablation Studies

We conducted ablation studies on the SceneFlow dataset with the following settings: (1) Student Model trained on clean scenes only (*Student-C*). (2) Student Model trained on foggy scenes only (*Student-F*). (3) Student Model trained on clean-and-fog mixed data (*Student-Mix*). For the student model, we use a simplified version of [9] as the baseline.

As shown in Table I, Training on clean or foggy images independently yields favorable results within the respective image domains. However, a significant decline in performance is observed when transitioning to a different image domain, as exemplified by *Student-C* (Clean \rightarrow Foggy: 0.68 \rightarrow 8.69) and *Student-F* (Foggy \rightarrow Clean: 0.580 \rightarrow 1.171). Employing mixed training, as demonstrated by *Student-Mix*, can partially mitigate this issue, but it is still inferior to the best performance in each image domain. The *Teacher Model* using adaptively fused clean and foggy features substantially outperforms other models, achieving an EPE of 0.352. However, its prerequisite for dual inputs is impractical for inference in real-world applications. With distillation from the *Teacher Model*, our *Student+Dist* using *Student Model* for inference demonstrates superior results in both

TABLE I: Ablation studies on the SceneFlow dataset. We report the EPE and P1-Error for both clean scenes and foggy scenes. 'C' means training on clean images only, while 'F' means training on foggy images only. 'Mix' denotes using mixed clean and foggy images for training. The 'Dist' is short for distillation, and 'Cont' is short for contrastive learning. Note that the *Teacher Model* requires both clean and corresponding foggy images for disparity estimation, which cannot be directly compared with other settings, marked by "*" in the table.

Method	Training Data	Loss			EPE↓		P1(%)↓	
		Disp Loss	Triplet Loss	Distillation Loss	Clean	Foggy	Clean	Foggy
Student-C	Clean	✓			0.628	8.69	7.3%	19.7%
Student-F	Foggy	✓			1.710	0.580	13.7%	8.4%
Student-Mix	Clean + Foggy	✓			0.731	0.560	8.3%	8.1%
<i>Teacher Model</i>	<i>Clean + Foggy</i>	✓			0.352*		5.6%*	
Student + Dist	Clean + Foggy	✓		✓	0.677	0.499	7.5%	7.2%
Student + Cont	Clean + Foggy	✓	✓		0.613	0.449	7.0%	6.9%
Student + Dist + Cont	Clean + Foggy	✓	✓	✓	0.552	0.401	6.5%	6.1%

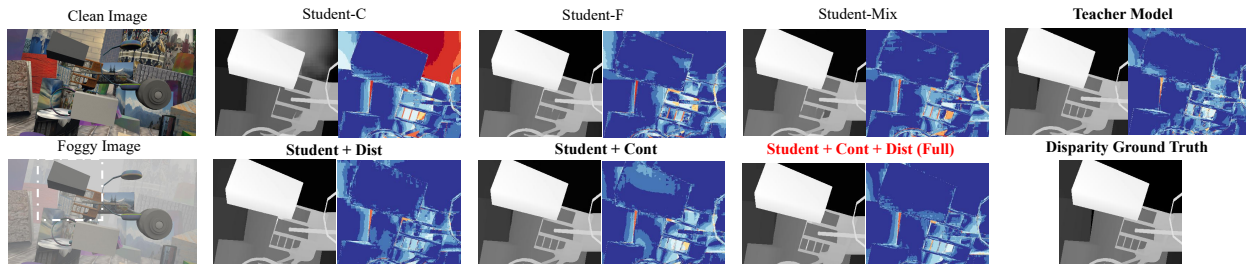


Fig. 4: Visualizations of ablation study on SceneFlow dataset. The first column is the predicted disparity for each group, while the second is the corresponding error map. We enlarged the selected part (white bounding box) for easier viewing.

TABLE II: Quantitative comparison of *CFDNet* and other methods on the SceneFlow dataset. We compared the results on clean and foggy data. "*" represents our reimplementation results. **Bold: best.**

Testing	Metric	Standard Stereo	Sequential		Physics-Based Joint-Learning			<i>CFDNet</i> (Ours)
		RaftStereo [9]	FFANet* [30] + RaftStereo	SDNet* [24]	SSMDNet* [8]	FoggyStereo* [7]		
Clean	EPE↓	0.69	0.76	4.26	3.61	1.00	0.55	
	3px(%)↓	3.7%	3.8%	24.3%	19.8%	4.8%	3.5%	
Foggy	EPE↓	1.87	1.33	2.48	2.08	0.66	0.40	
	3px(%)↓	11.6%	7.9%	16.3%	13.8%	4.0%	2.9%	

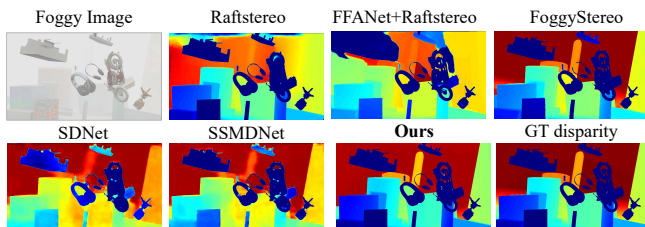


Fig. 5: Qualitative comparison on SceneFlow dataset foggy images with other superior works.

clean and foggy scenes compared to *Student-Mix*. Moreover, the triplet loss for contrastive learning embedded in *Student+Cont* further boosts robustness across both domains. Combining feature distillation and contrastive learning as our *CFDNet* framework, we observed peak performance, showing a 24.4% and 28.3% improvement over *Student* for clean and foggy scenes, respectively. As illustrated in Figure 4, our proposed network, incorporating contrastive learning feature distillation, exhibits superior performance by producing more structural details and fewer errors, while the baseline settings struggle to produce accurate estimations.

F. Performance Evaluation

Compared Methods: We compare our proposed *CFDNet* (*Student Model* with full loss functions) with three other methods, including standard stereo matching: RaftStereo [9], sequential two-stage dehazing and stereo matching: FFANet* [30]+RaftStereo, and Physics-based joint dehazing and stereo matching: SDNet [24], SSMDNet [8] and FoggyStereo [7]. Since the SDNet, SSMDNet, and FoggyStereo's codes were unavailable, we reimplemented them with the original settings and trained them on the same data. We conducted experiments on the famous SceneFlow and KITTI datasets with synthetic foggy images as well as the PixelAccurateDepth dataset with real foggy scenes for generation evaluation.

SceneFlow: As shown in Table II, reimplemented methods achieved competitive or superior performance in foggy images compared to the number in original papers. Our proposed *CFDNet* achieves at least 39.4% improvement in the foggy scenes compared with the second-best FoggyStereo, and a 78.6% improvement compared with the RAFT-Stereo trained on clean scenes. Furthermore, our method demonstrates superior performance on clean scenes with

TABLE III: The comparisons on the KITTI 2012 dataset. ‘*’ represents our reimplementations. **Bold: best.**

Method		Clean		Foggy	
		EPE↓ 3px(%) ↓		EPE↓ 3px(%) ↓	
Standard Stereo	RaftStereo	0.79	2.5%	0.84	3.0%
Sequential	FFANet*+RaftStereo	0.79	2.6%	0.83	2.9%
Physics-Based	SDNet*	1.66	10.0%	1.53	9.2%
	SSMDNet*	1.61	9.4%	1.51	9.0%
Joint-Learning	FoggyStereo*	0.78	2.7%	0.75	2.6%
<i>CFDNet</i> (Ours)		0.75	2.5%	0.70	2.1%

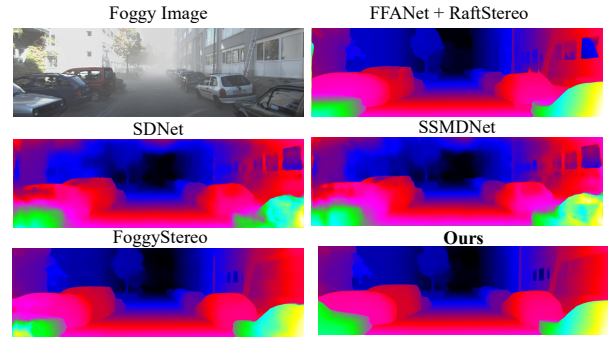


Fig. 6: Visualization comparison on the KITTI 2012 dataset in the foggy scenes with other superior works.

TABLE IV: The generalization evaluation on real-world clean and foggy scenes of PixelAccurateDepth Dataset with KITTI pre-trained model. ‘*’ represents our reimplementations. **Bold: Best.**

Method		Clean						Foggy(45m)					
		RMSE↓	MAE↓	SRD↓	ARD↓	SILog↓	$\delta_3 \uparrow$	RMSE↓	MAE↓	SRD↓	ARD↓	SILog↓	$\delta_3 \uparrow$
Standard Stereo	RaftStereo	1.57	0.65	0.13	4.02	9.23	99.80%	2.50	1.34	0.36	8.57	13.79	99.05%
Sequential	FFANet+RaftStereo	1.68	0.68	0.14	4.10	9.86	99.71%	3.15	1.67	0.62	9.88	16.91	95.53%
Physics-Based	SDNet*	1.58	0.81	0.14	5.98	10.21	99.83%	2.67	1.68	0.53	13.01	18.72	98.28%
	SSMDNet*	1.61	0.82	0.15	6.10	10.66	99.81%	2.60	1.59	0.49	12.30	17.84	98.37%
Joint-Learning	FoggyStereo*	2.27	0.78	0.18	6.57	15.40	99.11%	2.80	1.29	0.76	10.09	18.23	98.91%
<i>CFDNet</i> (Ours)		1.46	0.59	0.12	3.87	8.99	99.82%	2.33	1.18	0.31	7.81	14.05	99.69%

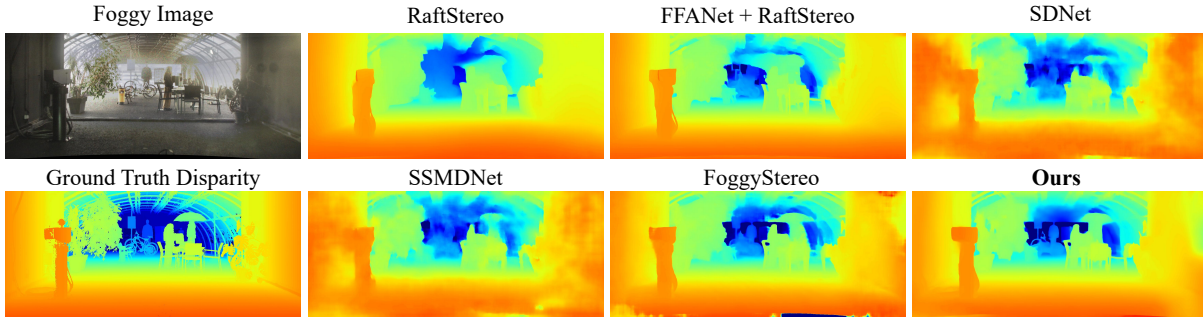


Fig. 7: Generalization visualization comparison on PixelAccurateDepth dataset foggy scenes with other superior works.

an EPE of 0.55, which indicates the effectiveness of our contrastive feature distillation framework for robust cross-domain performance.

KITTI 2012: As depicted in Table III, our proposed *CFDNet* outperforms existing methods when applied to foggy scenes. Additionally, our method exhibits competitive performance in clean scenes, even when compared to RaftStereo trained solely on clean images. This showcases the robustness of our contrastive feature distillation framework. Fig. 6 provides visual evidence of *CFDNet*’s proficiency in generating clear and consistent disparity maps, even when operating in foggy autonomous driving scenarios.

PixelAccurateDepth: As shown in Table IV, our proposed *CFDNet* demonstrates outstanding generalization performance on both clean scenes and foggy scenes except for the SDNet [24] with the δ_3 metric in the clean scenes. Moreover, our method significantly outperforms existing physics-based joint dehazing and stereo-matching approaches, particularly in metrics like SRD and ARD. The visualizations in Fig. 7 offer compelling evidence of *CFDNet*’s capacity for reliable

disparity estimation even in foggy conditions.

V. CONCLUSIONS

In this paper, we have presented the *CFDNet*, a novel stereo-matching network for both clean and foggy scenes. Diverging from previous methods that emphasize physical-based joint dehazing and stereo matching at the image level, our approach capitalizes on feature-level optimization using a contrastive learning incorporated feature distillation framework. This involves adaptive feature fusion for training the *Teacher Model* and feature distillation to boost the *Student Model*’s feature representation ability. Meanwhile, we implement a contrastive learning mechanism to balance the feature dependence on fog depth hints and clean matching features after distillation. Massive experiments on the SceneFlow, KITTI, and real-fog PixelAccurateDepth datasets have demonstrated the effectiveness of our proposed method in foggy scenes as well as our generalization ability on clean scenes.

REFERENCES

- [1] Y. Pang, J. Nie, J. Xie, J. Han, and X. Li, "Bidnet: Binocular image dehazing without explicit disparity estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5931–5940.
- [2] J. Nie, Y. Pang, J. Xie, J. Pan, and J. Han, "Stereo refinement dehazing network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3334–3345, 2021.
- [3] W. Treible, P. Saponaro, S. Sorensen, A. Kolagunda, M. O'Neal, B. Phelan, K. Sherbondy, and C. Kambhamettu, "Cats: A color and thermal stereo benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2961–2969.
- [4] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. IEEE, 1999, pp. 820–827.
- [5] R. T. Tan, "Visibility in bad weather from a single image," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.
- [6] B. Wróński, "Volumetric fog: Unified compute shader based solution to atmospheric scattering," in *ACM SIGGRAPH*, vol. 1, no. 2, 2014, p. 3.
- [7] C. Yao and L. Yu, "Foggystereo: Stereo matching with fog volume representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 043–13 052.
- [8] T. Song, Y. Kim, C. Oh, H. Jang, N. Ha, and K. Sohn, "Simultaneous deep stereo matching and dehazing with feature attention," *International Journal of Computer Vision*, vol. 128, pp. 799–817, 2020.
- [9] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.
- [10] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048.
- [11] A. Geiger, P. Lenz, C. Stillér, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [12] T. Gruber, M. Bijelic, F. Heide, W. Ritter, and K. Dietmayer, "Pixel-accurate depth evaluation in realistic driving scenarios," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 95–105.
- [13] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 650–656, 2006.
- [14] H. Hirschmüller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 807–814.
- [15] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo-stereo matching with slanted support windows," in *Bmvc*, vol. 11, 2011, pp. 1–11.
- [16] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2. IEEE, 2001, pp. 508–515.
- [17] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 7, pp. 787–800, 2003.
- [18] J. Chang and Y. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.
- [19] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [20] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3268–3277.
- [21] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 919–21 928.
- [22] T. Gruber, F. Julca-Aguilar, M. Bijelic, and F. Heide, "Gated2depth: Real-time dense lidar from gated images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1506–1516.
- [23] L. Caraffa and J.-P. Tarel, "Stereo reconstruction and contrast restoration in daytime fog," in *Asian conference on computer vision*. Springer, 2012, pp. 13–25.
- [24] T. Song, Y. Kim, C. Oh, and K. Sohn, "Deep network for simultaneous stereo matching and dehazing," in *BMVC*, 2018, p. 5.
- [25] S. Negahdaripour and A. Sarafraz, "Improved stereo matching in scattering media by incorporating a backscatter cue," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5743–5755, 2014.
- [26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [27] M. Hong, Y. Xie, C. Li, and Y. Qu, "Distilling image dehazing with heterogeneous task imitation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3462–3471.
- [28] S. Lee, T. Son, and S. Kwak, "Fifo: Learning fog-invariant features for foggy scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 911–18 921.
- [29] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.
- [30] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 908–11 915.
- [31] Z. Chen, Z. He, and Z.-M. Lu, "Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention," *arXiv preprint arXiv:2301.04805*, 2023.
- [32] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.
- [33] J. H. Lambert, *Photometria sive de mensura et gradibus luminis, colorum et umbrae*. sumptibus viduae E. Klett, typis CP Detleffsen, 1760.
- [34] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 158–22 169, 2020.