Variational Simultaneous Stereo Matching and Defogging in Low Visibility

Yining Ding[1]
yd2007@hw.ac.uk
Andrew M. Wallace[1]
a.m.wallace@hw.ac.uk
Sen Wang[2]
sen.wang@imperial.ac.uk

[1] Edinburgh Centre for Robotics
Heriot-Watt University
Edinburgh, UK

[2] Department of Electrical and Electronic Engineering & Imperial-X
Imperial College London
London, UK

Abstract

Given a stereo pair of daytime foggy[1] images, we seek to estimate a dense disparity map and to restore a fog-free image simultaneously. Such tasks remain extremely challenging in low visibility, partially preventing modern autonomous vehicles from operating safely. In this paper, we propose a novel simultaneous stereo matching and defogging algorithm based on variational continuous optimisation. It effectively fuses depth cues from disparity and scattering to achieve accurate depth estimation as the first step. Then the depth information is used to help restore a defogged image by leveraging a photo-inconsistency check. Extensive experiments on both synthetic and real data show the proposed algorithm outperforms comparative methods in all metrics on depth estimation, and produces visually more appealing defogged images.

1 Introduction

Dense and accurate depth estimation is essential for autonomous vehicles. Combined with a corresponding high-fidelity intensity image, depth information can benefit high-level vision tasks such as object detection [40] and semantic segmentation [44]. In comparison with active sensors, such as LiDAR and radar [19], video cameras are ubiquitous and cost-effective, and can infer scene depths from disparity provided the correspondence problem can be solved. Further, clear intensity images aid object recognition and help human drivers plan and act safely. However, like LiDAR, video camera perception degrades in adverse weather conditions, such as fog and snow. Whereas radar systems operate well in such conditions, they have relatively poor resolution and are not well interpreted by a human driver.

To solve these problems we propose simultaneous stereo depth reconstruction and defogging, where video cameras suffer from image colour shift and reduced local contrast. Existing stereo matching algorithms are predominantly developed under the assumption of clear scenes. Meanwhile, the vast majority of the literature on defogging addresses single

[1] In this work we do not distinguish between fog and haze because they are caused by similar atmospheric particles and the transition between them is gradual [37]. We focus on thick fog where the visibility is ≤ 40 meters.

images. There is very little work that tackles these two tasks simultaneously, even though they are deeply linked by scene depth, which can be inferred from the disparity [43] of stereo matching and scattering [13] of the fog model respectively. We expect that both results can be improved by better exploiting this underlying connection.

We propose a novel algorithm within the framework of continuous optimisation which takes a stereo pair of foggy images to simultaneously estimate disparity and perform defogging. Our main contributions are *threefold*: a) we design an anisotropic weighting scheme to allow for non-uniform penalty parameters which are seamlessly incorporated in the disparity optimisation process; b) we propose a customised regularisation term which effectively injects disparity cues from scattering by encouraging gradient alignment; c) we demonstrate, through extensive experiments in both synthetic and real scenes, that our method achieves very strong performance in both stereo matching and defogging compared with state-of-the-art (SOTA) methods, especially in extremely foggy scenarios. Our approach is based on variational methods that are easy to make parallel for acceleration. Moreover, it does not require training data containing foggy images with corresponding clear image and ground truth dense depth data. The acquisition of such data in *real outdoor* scenes is time-consuming at best and not always possible.

2 Related Work

2.1 Stereo Matching

The problem of stereo matching is to find visual correspondences between a pair of images, which can then be used to infer depths. Conventional methods, being either local (*e.g.* [20]) or global (*e.g.* [3, 23, 24]), usually comprise the following four steps [43]: matching cost computation, cost aggregation, disparity computation/optimisation, and disparity refinement. There has recently been a surge in deep learning based approaches [8, 11, 28], which demonstrate fruitful performance. However, the vast majority of the published work has been evaluated on clear scenes. Fog introduces complex visual effects. Experiments in [18] show the performance of some of the aforementioned methods degrade rapidly in presence of fog.

2.2 Defogging

The atmospheric scattering model, summarised in [33], states that the observed foggy image $I \in \mathbb{R}^{H\times W\times 3}$ is a convex combination of the latent clear image $J \in \mathbb{R}^{H\times W\times 3}$ and the atmospheric light $A \in \mathbb{R}^3$. The coefficients are controlled by a transmission map $t \in [0,1]^{H\times W}$ (Eq. (1)). Assuming homogeneous fog, t is determined by a constant scattering coefficient β and the distance d between a scene point and the camera (Eq. (2)). d can be further related to scene depth z given the camera intrinsic parameters. β encodes the fog density and is closely linked to visibility (*i.e.* the meteorological optical range [35, Chap. 9]) in meters (Eq. (3)).

$$I = Jt + A(1-t) \qquad (1) \qquad t = \mathrm{e}^{-\beta d} \qquad (2) \qquad vis = -\ln(0.05)/\beta \qquad (3)$$

Single Image Defogging This ill-posed problem amounts to the recovery of J given I only. Conventional methods [1, 9, 22] rely on some prior information on J or t, and mostly follow the same pipeline of first calculating A, then estimating and refining t, and finally inverting Eq. (1) to recover J. Recently, many deep learning based approaches have been proposed. Some pioneering work [6, 37] replaces only the t estimation stage by a convolutional neural

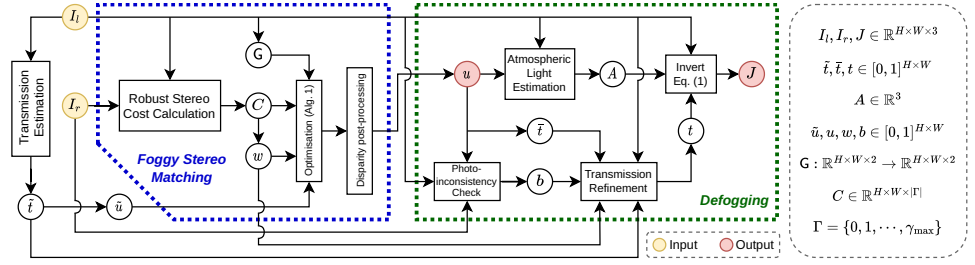


Figure 1: A block diagram of our method. The proposed two-stage system consists of a Foggy Stereo Matching module and a Defogging module. The former estimates a dense normalised disparity map u from a rectified stereo pair of foggy images $I_{\{l,r\}}$, then the latter performs defogging and restores a fog-free image J.

network. Later end-to-end methods [10, 25, 49] have treated single image defogging as an image-to-image regression problem.

Stereo Image Defogging This topic is closely linked to simultaneous stereo matching and defogging (see Sec. 2.3) but is inherently different in that no disparity map is estimated explicitly as a system output. Such an approach is justified by the observation that a small error in the estimated disparity may result in a large deviation in depth and thus also in the defogged image. In the light of this, various deep neural networks [34, 36] have been proposed. To prepare training data, most of them resort to completely synthetic scenes, *e.g.* [30], or add synthetic fog to real indoor images, *e.g.* [39]. Real, dense ground truth data, with and without fog, is very difficult to acquire. Some authors have added fog to intensity data from real outdoor scenes [26, 38], for which dense pseudo-ground truth depth data has been created by either monocular depth estimation [29] or stereoscopic inpainting [46]. Such a process can introduce undesirable artefacts in the synthesised foggy images.

2.3 Simultaneous Stereo Matching and Defogging

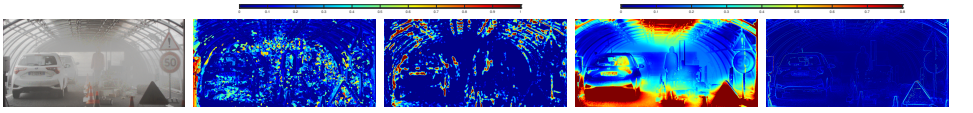
There has been a limited amount of existing work on simultaneous stereo matching and defogging. Early approaches [7, 27] recast the problem as energy minimisation of a Markov Random Field, solved by the α-expansion algorithm [3] or loopy belief propagation [16]. Unlike our approach, these discrete optimisation algorithms are difficult to make parallel. More recently, deep learning based approaches [31, 41, 42] have been adopted, but the same problem of synthesising realistic outdoor stereo foggy images still exists, particularly considering the challenges of collecting large-scale training foggy and clear images.

3 Method

In a nutshell, our two-stage system, depicted in Fig. 1, takes a rectified stereo pair of foggy images $I_{\{l,r\}}$ as input, and generates a dense normalised disparity map u after the first stage and a defogged image J after the second stage. Both u and J are in the left frame. We explain the key modules of our system in the rest of this section.

3.1 Estimation of Transmission Map

We estimate an initial transmission map $\tilde{t}$ ($\tilde{\cdot}$ denotes variables from defogging) from I_l by applying an existing single image defogging method ([22] is used in all our experiments).



(a) I_l (b) $|a^{(0)} - u_{\text{gt}}|$ (c) w (d) $|\tilde{u} - u_{\text{gt}}|$ (e) $|\nabla\tilde{u} - \nabla u_{\text{gt}}|$

Figure 2: Using a foggy scene whose left view is shown in (a), we pixel-wise plot: (b) the absolute error of the initial discrete disparity $a^{(0)}$; (c) the calculated weight array w; (d) the absolute error of the disparity from transmission $\tilde{u}$; and (e) the absolute error of the gradient of the disparity from transmission $\nabla\tilde{u}$. All disparity values are normalised between 0 and 1.

Once $\tilde{t}$ is estimated, the corresponding distance map $\tilde{d}$ can be calculated using Eq. (2) given the value of β, which is derived using Eq. (3) given a known visibility. In the case of an unknown visibility, considering a moving vehicle, it is possible to estimate β [21, 27]. However, it is difficult to do this from a single, static pair of images. Therefore, in our experiments β is assumed to be known. Given $\tilde{d}$ and the camera intrinsic parameters, the corresponding disparity map $\tilde{u}$ is calculated and fed into the subsequent foggy stereo matching block.

3.2 Foggy Stereo Matching

The foggy stereo matching block takes $I_{\{l,r\}}$ and $\tilde{u}$ as input, and estimates u as output. Our method is built upon [24], which introduces an auxiliary discrete variable a to decouple a convex regularisation term from a non-convex data term. The optimisation problem is then solved iteratively and alternately w.r.t. u and a (see [24] for full details). We propose two major extensions: a) to incorporate a weight array w which penalises the discrepancy between u and a non-uniformly at different pixel locations; b) to add a regularisation term based on $\tilde{u}$ which effectively deploys depth cues from scattering via gradient alignment.

Non-uniform penalty parameters Inspired by the idea of more general augmenting terms in ADMM [2], we incorporate a weight array $w \in [0,1]^{H\times W}$ to penalise the discrepancy between u and a non-uniformly at different pixel locations. The rationale for doing this is that the initial discrete disparity $a^{(0)}$, which is obtained from point-wise minimising a robust stereo matching cost volume $C \in \mathbb{R}^{H\times W\times|\Gamma|}$ (computed by first calculating the Hamming distance between the left and right Census Transforms [48] then locally aggregating the cost by the adaptive support-weight [47]) along its last dimension via exhaustive search, can be unreliable in certain (*e.g.* textureless) regions (see Fig. 2b). We want to penalise the difference between u and a to a lesser extent in such regions, but impose a larger penalty in regions where $a^{(0)}$ is more reliably estimated. To this end, at a pixel position (x,y) we empirically use the following soft step function (inspired by [1]) to generate the weight: $w_{x,y} = \min\{1,\ \max\{0.003,\ C^{\star\star}_{x,y}/C^{\star}_{x,y} - 1.15\}\}$, where $C^{\star}_{x,y}$ is the lowest cost (assume this occurs at disparity $\gamma^{\star}_{x,y}$) over the whole disparity range Γ, $C^{\star\star}_{x,y}$ is the lowest cost over Γ excluding disparities at $\{\gamma^{\star}_{x,y},\ \gamma^{\star}_{x,y}\pm 1,\ \gamma^{\star}_{x,y}\pm 2\}$. After calculating all entries of w, we normalise it so that $w \in [0,1]^{H\times W}$. Note in Fig. 2c that w has small values where errors of $a^{(0)}$ in Fig. 2b are large. We also apply $w_{x,y}$ to the per-pixel matching cost $C_{x,y} \in \mathbb{R}^{|\Gamma|}$ by scalar multiplication.

Depth cues from scattering by gradient alignment A naïve way of using $\tilde{u}$ is to consider it as a direct measurement of u and therefore create a data term. However, as single image defogging is severely ill-posed, $\tilde{t}$ may not be reliably estimated. Moreover, inaccurate β or inhomogeneous fog can also cause $\tilde{u}$ to contain large errors (see Fig. 2d). To overcome this issue we derive a regularisation term from $\tilde{u}$. More specifically, we encourage the non-zero gradient of the disparity to estimate u to occur at the same locations as the

non-zero gradient of $\tilde{u}$. Hence, we penalise inconsistency in non-zero gradient locations, as opposed to in gradient magnitude difference. A similar idea is used in [9] but to recover intensity images. Mathematically, we include a regularisation term $\|\nabla u - \nabla\tilde{u}\|_{1,0}$[2], where $\nabla : \mathbb{R}^{H\times W} \to \mathbb{R}^{H\times W\times 2}$ denotes a discrete gradient operator with Neumann boundary conditions. However, in practice the ℓ_0 minimisation problem is difficult to solve so we use the ℓ_1-norm $\|\nabla u - \nabla\tilde{u}\|_{1,1}$ instead, which can be rewritten as $\|\nabla(u-\tilde{u})\|_{1,1}$ due to the linearity of ∇. Note that the error in Fig. 2e is much smaller than in Fig. 2d. This term is applied with second-order Total Generalised Variation [4] (TGV^2, which promotes piece-wise planar surfaces) to constitute composite regularisors.

Optimisation With the above two modifications, we formulate the optimisation problem:

$$\underset{u,v}{\text{minimise}}\ \lambda_d \sum_{x,y} w_{x,y} C_{x,y}\left(u_{x,y}\right) + \lambda_s \left\|\mathsf{G}\left(\nabla u - v\right)\right\|_{2,1} + \lambda_a \left\|\boldsymbol{\nabla} v\right\|_{2,1} + \lambda_t \left\|\nabla(u-\tilde{u})\right\|_{1,1} + \iota_{[0,1]^{H\times W}}(u), \tag{4}$$

where $\mathsf{G} : \mathbb{R}^{H\times W\times 2} \to \mathbb{R}^{H\times W\times 2}$ denotes an anisotropic diffusion operator calculated from I_l, $v \in \mathbb{R}^{H\times W\times 2}$ denotes an additional variable to jointly optimise with u, $\boldsymbol{\nabla} : \mathbb{R}^{H\times W\times 2} \to \mathbb{R}^{H\times W\times 4}$ denotes a discrete gradient operator with Neumann boundary conditions, $\iota_{[0,1]^{H\times W}}(\cdot)$ is an indicator function to constrain the feasible set of u as it represents normalised disparities, and $\lambda_{\{d,s,a,t\}} \geq 0$ are tuning parameters.

We observe that all terms in Eq. (4), apart from the first one (*i.e.* the data term), are convex w.r.t. u. By introducing an auxiliary discrete variable a to decouple the convex terms from the non-convex data term, Eq. (4) can be recast as a constrained optimisation problem:

$$\begin{aligned}\underset{u,v,a}{\text{minimise}}\ & \lambda_d \sum_{x,y} w_{x,y} C_{x,y}\left(a_{x,y}\right) + \lambda_s \left\|\mathsf{G}\left(\nabla u - v\right)\right\|_{2,1} + \lambda_a \left\|\boldsymbol{\nabla} v\right\|_{2,1} \\ & + \lambda_t \left\|\nabla(u-\tilde{u})\right\|_{1,1} + \iota_{[0,1]^{H\times W}}(u), \quad \text{subject to} \quad u = a.\end{aligned} \tag{5}$$

We now form the augmented Lagrangian for Eq. (5):

$$\begin{aligned}& \lambda_d \sum_{x,y} w_{x,y} C_{x,y}\left(a_{x,y}\right) + \langle s, u-a\rangle + \frac{1}{2\theta}\sum_{x,y} w_{x,y}\left(u_{x,y}-a_{x,y}\right)^2 \\ & + \lambda_s \left\|\mathsf{G}\left(\nabla u - v\right)\right\|_{2,1} + \lambda_a \left\|\boldsymbol{\nabla} v\right\|_{2,1} + \lambda_t \left\|\nabla(u-\tilde{u})\right\|_{1,1} + \iota_{[0,1]^{H\times W}}(u),\end{aligned} \tag{6}$$

where $\langle\cdot,\cdot\rangle$ denotes inner product, s denotes the Lagrange multiplier, and $\theta > 0$ is a penalty parameter which controls how close u and a are drawn together globally.

Solver As [24] suggests, the optimisation problem of Eq. (6) can be iteratively solved by minimising it w.r.t. u (and v), minimising it w.r.t. a by point-wise exhaustive search, updating the Lagrange multiplier s, and finally decreasing θ to force u and a to be closer together. The above procedure is summarised in Algorithm 1. We use $\circ$ to denote the Hadamard product. The u minimisation problem (*i.e.* Line 3 in Algorithm 1) can be solved by the generalised Condat-Vu algorithm [12, 45]. See our supplemental material for full algorithmic details.

Disparity post-processing Some pixels in the leftmost region of the left frame cannot be seen by the right frame, causing their disparity values to be extremely close to zero after Algorithm 1. Since depth is inversely proportional to disparity, the depth errors are magnified, substantially impairing some of the depth error metrics. To overcome this issue we

[2]Throughout the paper we use $\|\cdot\|_{p,q}$ to denote a norm such that the p-norm is taken within the groups (*e.g.* across different colour channels) then the q-norm is taken between the groups (*e.g.* across different pixel locations). Using this notation, many commonly used sparsity-inducing functions can be conveniently yet unambiguously expressed, such as the anisotropic total variation ($p=1, q=1$) and the isotropic total variation ($p=2, q=1$).

Algorithm 1: The overall iterative algorithm to solve Eq. (6)

Input: C, w, $\tilde{u}$, G
Parameters : Γ, λ_d, λ_s, λ_a, λ_t, α, K
Output: $u^{(K+1)}$

Initialisation: $\forall x,y:\ u^{(0)}_{x,y} = a^{(0)}_{x,y} = \underset{a_{x,y}\in\Gamma}{\operatorname{argmin}}\ C_{x,y}(a_{x,y})$, $s^{(0)} = 0^{H\times W}$, $\theta^{(0)} = 1$;

for $k = 0, 1, \ldots, K$ **do**

$u^{(k+1)} = \underset{u}{\operatorname{argmin}} \langle s^{(k)}, u - a^{(k)}\rangle + \frac{1}{2\theta^{(k)}} \sum_{x,y} w_{x,y}\left(u_{x,y} - a^{(k)}_{x,y}\right)^2 + \lambda_s \left\|\mathsf{G}(\nabla u - v)\right\|_{2,1} + \lambda_a \left\|\nabla v\right\|_{2,1} + \lambda_t \left\|\nabla(u-\tilde{u})\right\|_{1,1} + \iota_{[0,1]^{H\times W}}(u)$;

$\forall (x,y):\ a^{(k+1)}_{x,y} = \underset{a_{x,y}\in\Gamma}{\operatorname{argmin}}\ \lambda_d w_{x,y} C_{x,y}(a_{x,y}) + s^{(k)}_{x,y}\left(u^{(k+1)}_{x,y} - a_{x,y}\right) + \frac{1}{2\theta^{(k)}} w_{x,y}\left(u^{(k+1)}_{x,y} - a_{x,y}\right)^2$;

$s^{(k+1)} = s^{(k)} + \frac{1}{2\theta^{(k)}} w \circ \left(u^{(k+1)} - a^{(k+1)}\right)$;

$\theta^{(k+1)} = \theta^{(k)}(1 - \alpha k)$;

return $u^{(K+1)}$

add a simple post-processing step (detailed in our supplementary material) to the output of Algorithm 1. It is worth mentioning that this affects disparities in the leftmost region *only*, as can be seen later in Fig. 6.

3.3 Defogging

Once we have established a dense disparity map u after the foggy stereo matching stage, we estimate the atmospheric light A. This is simply done by first locating where the median of the top 0.1% pixels with the smallest disparity values occurs, then selecting the intensity values of I_l at that pixel location as A. Next, assuming β is known, we can calculate a transmission $\bar{t}$ from u. However, the defogged image J may exhibit strong artefacts if $\bar{t}$ is directly used to invert Eq. (1). This can be attributed to factors, such as errors in u (*e.g.* caused by occlusion) and an inaccurate value of β. To overcome the issue we propose the following two extra steps.

Photo-inconsistency check We generate a weight array $b \in [0,1]^{H\times W}$ which encodes the photo-inconsistency between I_l and I_r given the disparity map u. b is computed by first warping I_r using u then measuring the Euclidean distance between the warped image and I_l in the CIELab colour space. The maximum distance is capped at a parameter ε. Finally we normalise b so all its values are between 0 and 1.

Transmission refinement We perform a transmission refinement to effectively fuse $\bar{t}$ with $\tilde{t}$. The former, which comes from the disparity u, is trustworthy in regions where the stereo weight w is high and photo-inconsistency b is low. The latter, which is derived from an initial transmission estimation step and irrelevant to the value of β, complements the former. In addition, we include a smoothness constraint to encourage discontinuities in our target transmission to align to I_l. Mathematically, we minimise the following quadratic form:

$$\underset{\mathbf{t}}{\text{minimise}}\ (\mathbf{t}-\bar{\mathbf{t}})^\intercal \bar{D}(\mathbf{t}-\bar{\mathbf{t}}) + (\mathbf{t}-\tilde{\mathbf{t}})^\intercal \tilde{D}(\mathbf{t}-\tilde{\mathbf{t}}) + \mu t^\intercal L t, \tag{7}$$

where $\bar{\mathbf{t}}$ and $\tilde{\mathbf{t}}$ are the column-major vectorisation of $\bar{t}$ and $\tilde{t}$ respectively, $\bar{D}$ and $\tilde{D}$ are diagonal matrices whose diagonal entries are from $\bar{d} = w \circ (1-b)$ and $\tilde{d} = (1-w)\circ b$ respectively, L is a five-point spatially inhomogeneous Laplacian matrix [15] derived from I_l, and $\mu > 0$ is a tuning parameter. Eq. (7) has a closed-form solution. Once Eq. (7) is solved, we reshape $\mathbf{t}$ to the image size and use it as the transmission map to invert Eq. (1) and recover J.

4 Experiments

We use the following two datasets for evaluation: the *synthetic* Virtual KITTI 2 dataset [5] (VKITTI2) and the *real* Pixel-Accurate Depth dataset [18] (PAD). See supplementary material for how we generate/select evaluation data, parameter setting and implementation[3].

For *defogging* evaluation, we compare our method with others that either have been extensively used as strong baselines [1, 6, 9, 22, 37], or represent the SOTA [10, 49]. To evaluate the deep learning based approaches [6, 10, 37, 49], we use their released, pre-trained models. For quantitative evaluation, we compute SSIM and PSNR values.

For *stereo matching* evaluation, we compare our method with [23, 24] (conventional), [8] (deep learning based) and [28] (SOTA). These baseline methods are applied directly to the foggy stereo pair as well as to the defogged stereo pair, the left and right frames of which are independently generated using the *best* (shown in *italics* in Tab. 1a, *i.e.* [22], and Tab. 2a, *i.e.* [10]) single image defogging methods. We apply the same disparity post-processing to [24] as a final step. To evaluate the deep learning approaches [8, 28], we use their released, pre-trained models on KITTI 2015. If a method generates disparities that are not 100% dense, background interpolation (as per the KITTI 2015 stereo benchmark) is performed. We adopt the D1-all disparity error and several depth error metrics from various KITTI benchmarks [17] and [14].

4.1 Virtual KITTI 2

Quantitative results are presented in Tab. 1 and qualitative results of two scenes are presented in Fig. 3. From Tab. 1b, the stereo matching performance of most baseline methods is significantly improved if defogging is applied beforehand (the only exception is [24] so in Fig. 3b we show its results *without* the defogging step). This is in line with the observation from Fig. 3a that [22] removes fog reasonably well at close and medium ranges but fails to reveal distant objects. Consequently, most baseline stereo matching methods that operate on the defogged image pair are not able to resolve objects at far range either (see the first two rows of Fig. 3b).

In addition, we investigate how the disparity results of all stereo matching methods vary if the input images get flipped *upside down*. This is illustrated in the last row of Fig. 3b, overlaid with the D1-all metric value change. The motivation behind this experiment is to determine to what extent stereo reconstruction is based on context information (*e.g.* the depth values in the foreground road are entirely predictable in VKITTI2 data without stereo) and to what extent dependent more generally on object geometry and reflectance. The key observation is that [24] and our method are able to produce very consistent results, whereas other methods witness noticeable performance degradation.

(a) Defogging

Method	SSIM	PSNR
DCP [code]	0.606	9.244
GRM [code]	0.458	7.993
NLD [code]	0.480	7.323
DehazeNet [code]	0.487	7.386
MSCNN [code]	0.499	7.537
PSD [code]	0.424	6.591
4KDehazing [code]	0.533	11.135
Ours	**0.668**	**11.248**

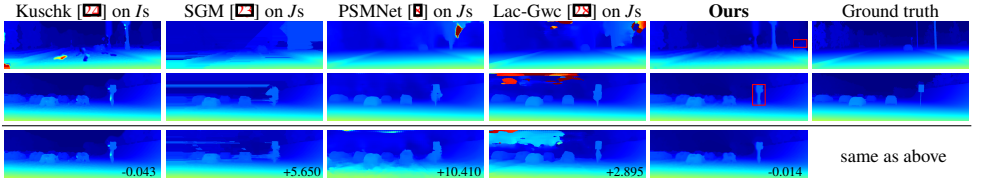
(b) Stereo matching

Method	D1-all	RMSE	MAE	SILog	Sq Rel	Abs Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
on *Is*									
Kuschk [code]	10.309	26.375	7.491	28.905	4.655	10.317	85.525	91.063	94.468
SGM [code]	33.061	35.845	13.578	53.615	12.087	27.524	71.718	79.749	84.294
PSMNet [code]	35.306	39.603	15.932	53.176	11.473	20.960	67.398	77.740	83.331
Lac-GwcNet [code]	24.026	38.028	14.218	52.408	10.448	16.331	76.025	82.039	85.542
on *DCP Js*									
Kuschk [code]	10.873	26.596	7.628	30.336	4.733	10.554	84.947	90.548	94.027
SGM [code]	23.813	32.590	11.242	43.379	9.813	21.567	78.282	84.656	88.338
PSMNet [code]	24.373	37.554	14.188	46.226	9.783	17.699	73.802	83.827	87.634
Lac-GwcNet [code]	15.309	34.753	11.632	42.382	8.180	11.894	83.390	87.771	89.922
Ours	**8.861**	**26.143**	**7.310**	**25.517**	**4.197**	**8.795**	**86.922**	**92.415**	**95.523**

Table 1: Quantitative results on VKITTI2. Ours performs the **best** in all metrics.

[3]Our code is publicly available at: https://github.com/tedyiningding/VSSMD.

(a) Defogging results. Our method is the best at revealing distant vehicles (blue rectangles), whereas all baselines methods fail to do so. It is also worth noting that although [49] outperforms other baseline methods in the averaged PSNR by a comfortable margin, its results are visually much less favourable (blurring and ghost objects) and therefore we consider [22] as the best baseline method.



(b) Disparity results. The first two rows show our method is the best at resolving small and distant objects (red rectangles) with the least overall visual artefacts. The last row shows [24] and our method are able to produce very consistent results after flipping the second row's input images upside down.

Figure 3: Qualitative results on VKITTI2

4.2 Pixel-Accurate Depth Dataset

Quantitative results are presented in Tab. 2 and qualitative results of three scenes covering the whole visibility range of interest are presented in Fig. 5. In Tab. 2b, we see that this time a prior defogging step does not improve but rather impairs the stereo matching performance of all baseline methods. Therefore, the disparity results of baseline methods presented in Fig. 5b directly use foggy stereo pairs as input.

In addition, we investigate how the performance of all stereo matching methods varies with visibility. In Fig. 4 we see that our method clearly excels in the depth RMSE for 20m visibility (4.5% better than [23] and about 26.9% better than [8, 24, 28]). A higher visibility gradually narrows down the performance gap between ours and [8, 28], which is in line with what can be observed from Fig. 5b.

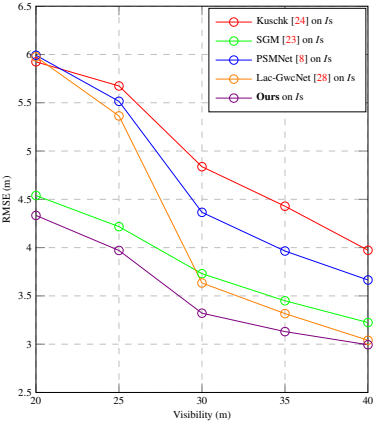


Figure 4: Depth RMSE vs. visibility on PAD

4.3 Ablation Study

We conduct an ablation study on the PAD dataset to better understand how different modules in our system contribute to defogging and stereo matching performance. Quantitative results

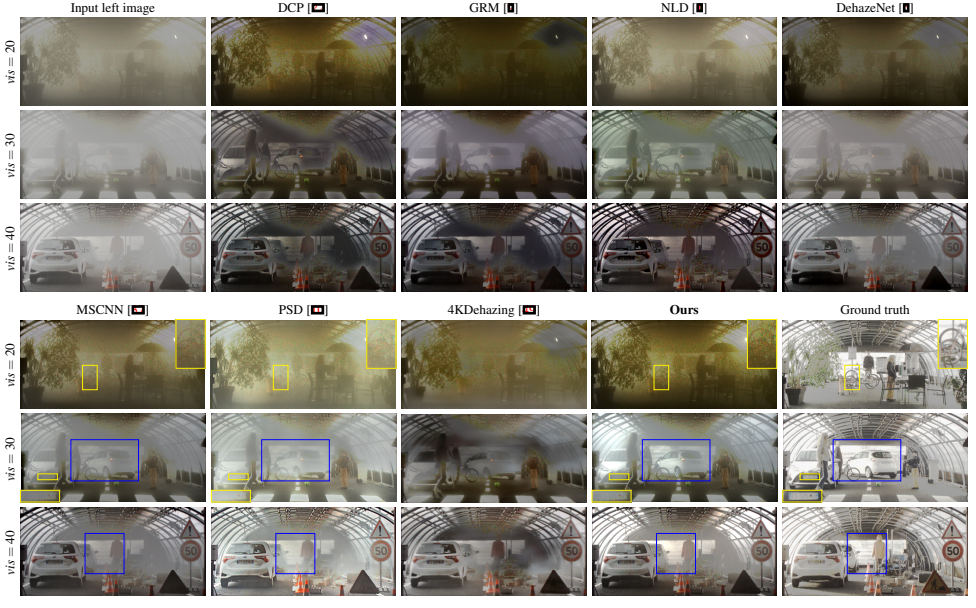
(a) Defogging

Method	SSIM	PSNR
DCP [11]	0.380	8.523
GRM [8]	0.404	8.732
NLD [8]	0.496	11.806
DehazeNet [8]	0.473	11.062
MSCNN [28]	0.519	12.023
PSD [23]	**0.588**	**16.676**
4KDehazing [29]	0.510	10.419
Ours	0.519	13.433
Ours ($w = 1$)	0.384	10.411
Ours ($\lambda_t = 0$)	0.526	13.458
Ours (w/o pp)	0.519	13.432
Ours (w/o tr)	0.352	10.072

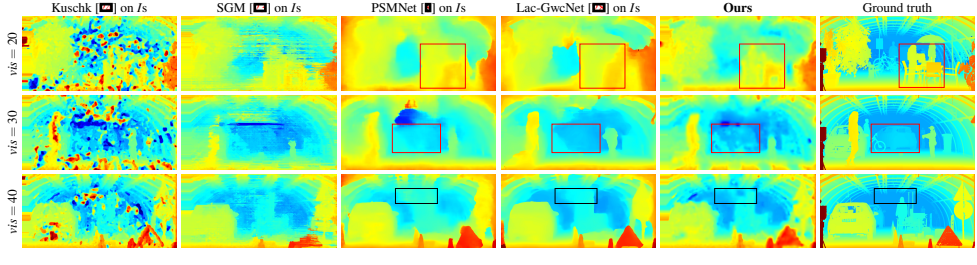
(b) Stereo matching

	Method	D1-all	RMSE	MAE	SILog	Sq Rel	Abs Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
on *Is*	Kuschk [11]	49.988	4.967	3.208	28.496	1.320	18.000	70.961	85.803	94.161
	SGM [28]	42.515	3.832	2.518	21.377	0.786	14.191	78.496	92.250	98.464
	PSMNet [8]	49.913	4.700	3.115	21.950	1.029	15.519	71.102	87.698	96.764
	Lac-GwcNet [29]	38.637	4.266	2.671	22.084	0.921	13.563	76.881	89.227	95.653
on *PSD Is*	Kuschk [11]	52.449	5.131	3.345	29.669	1.440	18.994	69.346	84.966	93.681
	SGM [28]	43.365	3.896	2.549	21.900	0.820	14.476	77.774	91.713	98.298
	PSMNet [8]	54.165	5.084	3.430	24.182	1.224	17.136	67.514	85.350	95.095
	Lac-GwcNet [29]	41.742	4.498	2.847	22.603	0.991	14.329	74.606	87.744	95.123
Ours		**37.647**	**3.550**	**2.282**	**19.078**	**0.651**	**12.352**	**80.435**	**93.959**	**98.871**
Ours ($w = 1$)		48.915	4.840	3.114	27.626	1.250	17.435	71.975	86.553	94.639
Ours ($\lambda_t = 0$)		42.765	4.233	2.701	22.636	0.893	14.476	75.953	89.987	97.739
Ours (w/o pp)		38.208	4.105	2.507	23.172	1.137	14.836	79.539	93.107	97.697

Table 2: Quantitative results on PAD. Ours performs the **best**/second best in all metrics.



(a) Defogging results. Compared to the *top two* baseline methods [11, 37], ours is better at removing fog from distant objects (blue rectangles) and preserving fine details (close-up of yellow rectangles). Note that our defogged images are visually more appealing despite trailing behind [11] in metric values.



(b) Disparity results. [24]'s results contain a large number of outliers. [23]'s results exhibit strong streaking artefacts. Our method preserves fine structures to a greater extent in extremely low visibilities (red rectangles), whereas the two deep learning based approaches [8, 28] fail to do so. In slightly better visibility (third row), [8, 28] start surpassing our method visually with sharper object edges at close range, but still suffer from blob artefacts at distant objects (black rectangles).

Figure 5: Qualitative results on PAD

are shown in the last four rows of Tab. 2a and the last three rows of Tab. 2b. Qualitative results are illustrated in Fig. 6. We use $w = 1$ to denote the case in which a uniform weight array (*i.e.* all ones) is used, $\lambda_t = 0$ to denote the case in which no depth cues from scattering are used, w/o pp to denote the case without performing the disparity post-processing, and w/o tr to denote the case without transmission refinement (*i.e.* $\bar{t}$ is directly used to invert Eq. (1) and recover J). We observe: a) by adopting the proposed non-uniform weighting scheme, both stereo matching and defogging performances are significantly improved; b) by employing depth cues from scattering, the stereo matching performance is moderately improved but there is not much impact on the defogging results; c) the disparity post-processing step greatly improves some depth error metrics such as SILog; we would like to point out that our method demonstrates very competitive stereo matching performance in most metrics even *without* the disparity post-processing; d) without the transmission refinement step, both defogging metrics become worse.

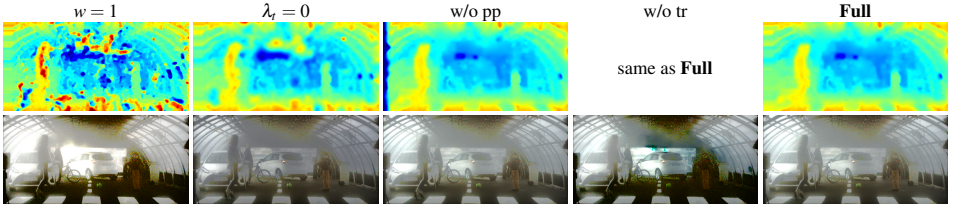


Figure 6: Qualitative results of ablation study on PAD

5 Conclusion

We have presented an approach within the framework of variational continuous optimisation that addresses the problem of simultaneous stereo matching and defogging in low visibility. As opposed to sequentially performing defogging then stereo matching, we directly use foggy images as the stereo matching input. Combining a depth error informed non-uniform weighting scheme with an effective way of extracting depth cues from scattering via gradient alignment enables us to reconstruct accurate disparity maps. The depth information is then used to assist image defogging. In comparison with methods based on deep learning, we do not require a training dataset which in any event cannot generally be acquired in the wild, as corresponding clear and degraded images cannot be acquired on the same scenes with moving actors. Evaluated on both synthetic and real datasets, our method surpasses comparative methods in all depth estimation metrics (up to 26.9% reduction in RMSE) and produces visually more appealing defogged images, particularly in extremely poor visibilities. For future work we consider: a) using motion information embedded in consecutive frames and incorporating more matching constraints to improve depth estimation results; b) adopting a more sophisticated fog model (*e.g.* blurring and fog inhomogeneity) to better recover intensity images.

Acknowledgements

We thank João F. C. Mota for helpful discussions. We are grateful for EPSRC funding for the Centre for Doctoral Training in Robotics and Autonomous Systems EP/S023208/1.

References

[1] Dana Berman, Tali treibitz, and Shai Avidan. Non-local image dehazing. In *CVPR*, 2016.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 2011.

[3] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001.

[4] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 2010.

[5] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

[6] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *TIP*, 2016.

[7] Laurent Caraffa and Jean-Philippe Tarel. Stereo reconstruction and contrast restoration in daytime fog. In *ACCV*, 2012.

[8] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018.

[9] Chen Chen, Minh N. Do, and Jue Wang. Robust image and video dehazing with visual artifact suppression via gradient residual minimization. In *ECCV*, 2016.

[10] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *CVPR*, 2021.

[11] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *NeurIPS*, 2020.

[12] Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 2013.

[13] Fabio Cozman and Eric Krotkov. Depth from scattering. In *CVPR*, 1997.

[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

[15] Zeev Farbman, Raanan Fattal, Dani Lischinski, and Richard Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. *TOG*, 2008.

[16] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 2006.

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[18] Tobias Gruber, Mario Bijelic, Felix Heide, Werner Ritter, and Klaus Dietmayer. Pixel-accurate depth evaluation in realistic driving scenarios. In *3DV*, 2019.

[19] Juan Guerrero-Ibáñez, Sherali Zeadally, and Juan Contreras-Castillo. Sensor technologies for intelligent transportation systems. *Sensors*, 2018.

[20] Marsha Jo Hannah. *Computer Matching of Areas in Stereo Images*. PhD thesis, Stanford University, 1974.

[21] Nicolas Hautiére, Jean-Philippe Tarel, Jean Lavenant, and Didier Aubert. Automatic fog detection and estimation of visibility distance through use of an onboard camera. *Machine Vision and Applications*, 2006.

[22] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *PAMI*, 2010.

[23] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 2007.

[24] Georg Kuschk and Daniel Cremers. Fast and accurate large-scale stereo reconstruction using variational methods. In *ICCVW*, 2013.

[25] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, 2017.

[26] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 2018.

[27] Zhuwen Li, Ping Tan, Robby T. Tan, Danping Zou, Steven Zhiying Zhou, and Loong-Fah Cheong. Simultaneous video defogging and stereo reconstruction. In *CVPR*, 2015.

[28] Biyang Liu, Huimin Yu, and Yangqi Long. Local similarity pattern and cost self-reassembling for deep stereo matching networks. In *AAAI*, 2022.

[29] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *PAMI*, 2015.

[30] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[31] Jeong-Yun Na and Kuk-Jin Yoon. Stereo vision aided image dehazing using deep neural network. In *Proceedings of the 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild*, 2018.

[32] Srinivasa G. Narasimhan and Shree K. Nayar. Vision and the atmosphere. *IJCV*, 2002.

[33] Srinivasa G. Narasimhan and Shree K. Nayar. Contrast restoration of weather degraded images. *PAMI*, 2003.

[34] Jing Nie, Yanwei Pang, Jin Xie, Jing Pan, and Jungong Han. Stereo refinement dehazing network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[35] World Meteorological Organization. Measurement of meteorological variables, 2014. URL https://library.wmo.int/index.php?lvl=notice_display&id=19673#.Yzw6YTfMJhE. Last accessed 4 October 2022.

[36] Yanwei Pang, Jing Nie, Jin Xie, Jungong Han, and Xuelong Li. Bidnet: Binocular image dehazing without explicit disparity estimation. In *CVPR*, 2020.

[37] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, 2016.

[38] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 2018.

[39] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[40] Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *ECCV*, 2014.

[41] Taeyong Song, Youngjung Kim, Changjae Oh, and Kwanghoon Sohn. Deep network for simultaneous stereo matching and dehazing. In *BMVC*, 2018.

[42] Taeyong Song, Youngjung Kim, Changjae Oh, Hyunsung Jang, Namkoo Ha, and Kwanghoon Sohn. Simultaneous deep stereo matching and dehazing with feature attention. *IJCV*, 2020.

[43] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010.

[44] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019.

[45] Bằng Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 2013.

[46] Liang Wang, Hailin Jin, Ruigang Yang, and Minglun Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *CVPR*, 2008.

[47] Kuk-Jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. *PAMI*, 2006.

[48] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV*, 1994.

[49] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Xiaobin Hu, Tao Wang, Fenglong Song, and Xiuyi Jia. Ultra-high-definition image dehazing via multi-guided bilateral learning. In *CVPR*, 2021.