

Neural Markov Random Field for Stereo Matching

Tongfan Guan¹

tfguan@link.cuhk.edu.hk

Chen Wang²

chenw@sairlab.org

Yun-Hui Liu^{1*}

yhliu@mae.cuhk.edu.hk

¹The Chinese University of Hong Kong

²Spatial AI & Robotics Lab, University at Buffalo

Abstract

Stereo matching is a core task for many computer vision and robotics applications. Despite their dominance in traditional stereo methods, the hand-crafted Markov Random Field (MRF) models lack sufficient modeling accuracy compared to end-to-end deep models. While deep learning representations have greatly improved the unary terms of the MRF models, the overall accuracy is still severely limited by the hand-crafted pairwise terms and message passing. To address these issues, we propose a neural MRF model, where both potential functions and message passing are designed using data-driven neural networks. Our fully data-driven model is built on the foundation of variational inference theory, to prevent convergence issues and retain stereo MRF’s graph inductive bias. To make the inference tractable and scale well to high-resolution images, we also propose a Disparity Proposal Network (DPN) to adaptively prune the search space of disparity. The proposed approach ranks 1st on both KITTI 2012 and 2015 leaderboards among all published methods while running faster than 100 ms. This approach significantly outperforms prior global methods, e.g., lowering D1 metric by more than 50% on KITTI 2015. In addition, our method exhibits strong cross-domain generalization and can recover sharp edges. The codes at <https://github.com/aeolusguan/NMRF>.

1. Introduction

Stereo matching is a critical component in computer vision, mimicking human binocular vision to cognize 3D information within the field of view [37]. Given a pair of images, it aims to determine the horizontal displacement of individual pixels in one image to align with the other. Stereo matching has been applied to many fields like 3D reconstruction [49], autonomous navigation [53], and augmented reality [59], bridging the gap between digital imagery and real-worlds.

Among the various methodologies employed in stereo matching, Markov Random Field (MRF) [50] stands out as one of the most widely used and effective models. MRFs

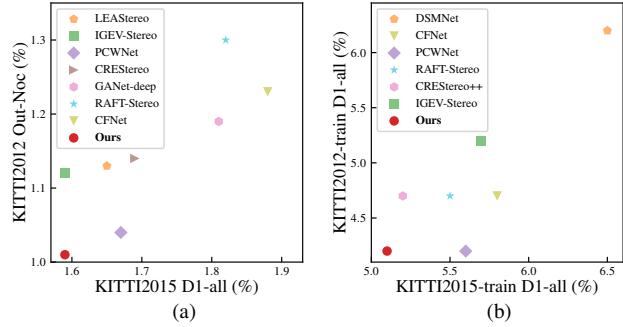


Figure 1. (a) Comparison with state-of-the-art stereo methods [11, 27, 30, 42, 43, 55, 61] on KITTI 2012 and 2015 leaderboards. (b) Cross-domain generalization comparison with current robust methods [21, 30, 42, 43, 55, 62]. All methods are only trained on the synthetic SceneFlow dataset [33] and evaluated on KITTI2012/2015 trainsets with fixed parameters.

leverage a probabilistic model to explain observed image features and enforce spatial coherence, producing piecewise smooth disparity maps. Due to their ability to reduce matching ambiguities in challenging regions [46], MRF and its variants [1, 20, 47, 60] have dominated the field before deep neural networks [22, 30, 58] emerged, according to Middlebury [37], a popular benchmark for stereo matching.

Although MRFs have achieved promising results, they are still often faced with difficulties because of hand-crafted *potential functions* and *message passing* procedures. Typically, an MRF’s potential function comprises a *unary* term that evaluates the similarity in intensity/gradient of matching pixels/features; and a *pairwise* term that penalizes solutions that violate certain spatial smoothness criteria [46]. However, such hand-crafted approaches cannot fully model all scenarios, e.g., carefully designed models for object occlusion may not be able to model abrupt changes in disparity at object boundaries. This inadequacy often results in an over-smoothed disparity map. Moreover, hand-crafted message passing may also struggle to handle complex pairwise relationships. This limitation can often lead to difficulties such as inaccurate disparity estimations or convergence issues. Despite the unary terms have been greatly improved by deep feature representations [8, 24, 25, 32, 44], the over-

*Corresponding author: Yun-Hui Liu

all accuracy of MRFs is still severely limited by the hand-crafted pairwise potential and message passing functions.

To tackle the above problems, we propose a Neural MRF (NMRF) model to learn the complicated pixel relationships in a data-driven manner, mitigating the inefficacy of manual design. The mean-field variational inference theory is leveraged to design neural modules that perform as unary/pairwise potential terms and message passing. Additionally, to make our NMRF tractable and scale well to high-resolution images, we propose a Disparity Proposal Network (DPN) which significantly prunes the search space of disparity with little sacrifice of performance. To the best of our knowledge, NMRF is the first fully data-driven stereo MRF model while retaining its strong graph inductive bias to handle uncertainty and ambiguity in image data.

Our NMRF model reports state-of-the-art accuracy on SceneFlow [33] and ranks 1st on KITTI 2012 [15] and 2015 [34] leaderboards among all published methods while running faster than 100 ms. Compared with previous global stereo networks, NMRF outperforms with a substantial margin, *e.g.*, reducing the *D1-bg* outlier ratio by more than 50% on KITTI 2015, 1.28% (Ours) *vs.* 2.85% (LBPS [25]). NMRF also exhibits state-of-the-art cross-domain generalization ability. When trained only on synthetic SceneFlow dataset, NMRF performs very well on real datasets KITTI [15, 34], Middlebury [38], and ETH3D [39]. Furthermore, NMRF is able to recover sharp depth boundaries, as shown in Fig. 2a, which is key to downstream tasks, such as 3D reconstruction and object detection.

In summary, our contributions include:

- We introduce a novel fully data-driven MRF model for stereo matching that can effectively learn complicated relationships between pixels from data.
- We develop a search space pruning module that largely reduces the computation load of neural MRF inference, which is also valuable in other dense matching tasks.
- Our architecture achieves state-of-the-art results on popular benchmarks in terms of both accuracy and robustness.

2. Related Work

Since MC-CNN [58], deep learning has been leveraged for unary matching cost computation [8, 19, 24, 32, 57, 58], cost volume aggregation [6, 10, 11, 17, 22, 42, 43, 54, 56, 61, 62], and iterative disparity refinement [5, 21–23, 27, 29, 30, 33, 48, 52, 55, 56]. Laga *et al.* [26] give a thoroughly survey on deep techniques for stereo matching. This section focuses on MRF-related stereo networks.

Stereo MRF networks. MRFs formulate stereo matching as a pixel-labeling problem, and assign every pixel o a disparity label z_o . The set of all pixel-label assignments is denoted by $\{z_o\}$. We write MRFs’ probabilistic model as:

$$p(\{z_o\}, \{\mathbf{x}_o\}) \propto \prod_o \Phi(z_o, \mathbf{x}_o) \prod_{(o,p)} \Psi(z_o, z_p), \quad (1)$$

where Φ and Ψ are non-negative unary and pairwise *potential functions*, respectively; (o, p) represents a pair of neighboring pixels; and \mathbf{x}_o is the observed pixel features. The Maximum A Posterior (MAP) estimate of $\{z_o\}$ is equivalent to energy minimization by taking *log* of Eq. (1).

Typically, Φ and Ψ are hand-crafted based on stereo domain knowledge, *e.g.*, intensity constancy and piecewise coplanar [2, 47]. The pioneering MC-CNN [58] generalizes manually designed unary potential Φ with a siamese network, which performs feature extraction from image patches and computes unary costs based on a fully-connected DNN. Chen *et al.* [8] achieved 100 \times speed-up compared to MC-CNN by replacing the fully-connected DNN with a dot product layer at the cost of little performance drop. Instead of independent predictions on pairs of image patches, Luo *et al.* [32] compared a patch in the left image with a horizontal stripe in the right image to extract marginal distributions over all possible disparities. However, these methods still leverage the hand-crafted pairwise potential and message passing of SGM [20]. Considering the difficulty of tuning SGM parameters (P_1, P_2) to accurately penalize disparity discontinuities for different cases, Seki *et al.* [41] trained a neural network to provide adaptive parameters $P_1(o)$ and $P_2(o)$ for every pixel o . Similarly, PBCP [40] set them based on the disparity confidence map estimated with a CNN. Hybrid CNN-CRF [24] performed feature matching on complete images, and this setting firstly enabled end-to-end joint learning of SGM message passing and the unary/pairwise CNNs. Moreover, GANet [61] proposed a differentiable approximation of SGM by replacing user-defined parameters with learned guidance weight matrix. Recently, LBPS [25] adapted Belief Propagation (BP) to learning formulation through a differentiable loss defined on marginal distributions, making graphical models fully compatible with deep learning. However, the application of hand-crafted message passing functions didn’t achieve top performance¹ due to its limited ability to handle complicated pairwise relationships.

Neural message passing. Gilmer *et al.* [16] demonstrated that neural message passing (NMP) is the cornerstone of recent successful models on graph-structured data. Specifically, it exchanges vector messages between nodes and updates node embeddings using neural networks. In computer vision tasks, it has been employed to exchange descriptor information among keypoints in sparse feature matching [36, 45] and solve the maximum network flow formulation of multiple object tracking [3]. Dai *et al.* [12] noted that neural message passing can be derived through certain embedded inference on probabilistic graphical models (PGM) [18]. This observation motivates us to build our

¹The overall accuracy of GANet [61] mainly owing to 3D convolutional cost aggregation. Approximate SGM aggregation alone didn’t achieve satisfactory accuracy, as evidenced by its ablation studies.

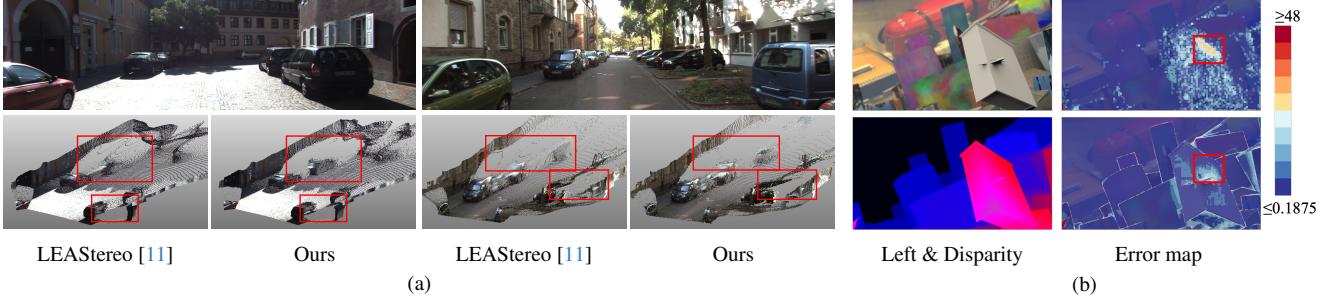


Figure 2. (a) Stereo point cloud comparison between LEAStereo [11] and our method on KITTI test set. Notice how our approach notably alleviates flying pixels near object boundaries, which is well-known as over-smoothing problem [7]. Please zoom in for more details. (b) Left column: left image (top) and disparity estimation (bottom), Right column: color-coded error map of pixelwise best proposal (top) and disparity estimation (bottom). This method even recovers from proposal failure (marked with red) in the large textureless region.

neural message passing function on the foundation of variational inference theory, to prevent convergence issues of MRF optimization. Given that Transformer [13, 31] is also a form of message passing technique, we incorporate key elements of Transformer into our message passing function while retaining the graph inductive bias of stereo MRF.

Label space pruning. Many algorithms devoted to pruning MRF label space due to tractability or efficiency reason. Menze *et al.* [35] pruned two-dimensional flow space using descriptor matching. To prune the disparity search space, DeepPruner [14] designed a differentiable PatchMatch-based pruner to predict a confidence range for each pixel. CFNet [42] and UCSNet [9] gradually narrowed down the disparity search space in the cascaded multi-scale manner, where next scale’s search range is generated based on the disparity estimate of current scale. The unimodal search range $[l_o, r_o]$ outputted by these popular methods [9, 14, 42] for each pixel o , however, may be susceptible to local optimum. In contrast, the proposed Disparity Proposal Network (DPN) has the ability to predict multi-modal proposals.

3. Methodology

Given an image pair (I^L, I^R) , we propose to estimate the disparity map via a Neural Markov Random Field (NMRF) model. An overview of the approach is presented in Fig. 3.

This approach consists of two stages. The first stage employs a NMRF model to infer disparities on the coarse level ($1/8$) features (Sec. 3.2). To make NMRF efficient, a Disparity Proposal Network is proposed to prune candidate space (Sec. 3.3). The next stage performs disparity refinement on the fine level ($1/4$) features (Sec. 3.4). A local feature CNN provides the coarse and fine level features for both stages.

Local feature CNN. To extract multi-level features from the pair of images, we employ a siamese convolutional network, which comprises a stack of residual blocks, instance normalization and downsampling layers. The coarse level features at $1/8$ of the original image resolution are denoted as \tilde{F}^L and \tilde{F}^R , while the fine-level features at $1/4$ of the original image resolution are denoted as \hat{F}^L and \hat{F}^R .

3.1. Neural MRF Formulation

We start by formulating the NMRF model for stereo matching before implementing it in Sec. 3.2. Given a discrete candidate label (disparity) set $L_o \subset \mathbb{R}^+$ for every pixel o , hand-crafted MRFs center on a carefully designed energy function over pixel-label assignments $\{z_o | z_o \in L_o\}$. However, the pairwise terms typically depend only on label difference $|z_o - z_p|$ while overlooking image content [46]. In order to incorporate additional information, such as 3D geometry and visual context, into potential functions, our NMRF represents each label with an *embedding* $\mathbf{z}_v \in \mathbb{R}^d$.

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ defines the discrete neural MRF as:

$$p(\{\mathbf{x}_v\}, \{\mathbf{z}_v\}) \propto \prod_{v \in \mathcal{V}} \Phi(\mathbf{x}_v, \mathbf{z}_v) \prod_{(u, v) \in \mathcal{E}} \Psi(\mathbf{z}_u, \mathbf{z}_v), \quad (2)$$

where each node v corresponds to a candidate label, and the edge (u, v) connects a pair of labels from neighbor pixels within a $M \times M$ image window. Intuitively, $\Phi(\mathbf{x}_v, \mathbf{z}_v)$ indicates the data likelihood of observed label feature \mathbf{x}_v given its latent embedding \mathbf{z}_v , while $\Psi(\cdot, \cdot)$ controls the penalty of pairwise label assignments. Given observed label features $\{\mathbf{x}_v\}$, our goal is to infer the latent embeddings $\{\mathbf{z}_v\}$, from which the disparity estimation will be deduced.

As shown in the 3rd part of Fig. 3, this graph has two types of edges. Intra-pixel edges, or *self* edges, $\mathcal{E}_{\text{self}}$, connect labels v to all other labels of the same pixel. Inter-pixel edges, or *neighbor* edges, $\mathcal{E}_{\text{neigh}}$, connect labels v to all labels of neighbor pixels. Self edges are usually overlooked in hand-crafted stereo MRFs. In contrast, this paper accounts for self edges either, but uses different potential functions, *i.e.*, Ψ_{neigh} and Ψ_{self} , for the two types of edges. Intuitively, Ψ_{neigh} allows compatible label pairs to support each other, whereas Ψ_{self} expects labels of the same pixel to compete with each other. Table 5 shows that this inductive bias for pairwise potential will bring considerable improvements.

In hand-crafted MRF settings, Φ and Ψ are typically designed based on stereo domain knowledge. In our neural MRF, however, we do not specify the exact form of potential functions Φ and Ψ . Instead, we will seek to implicitly

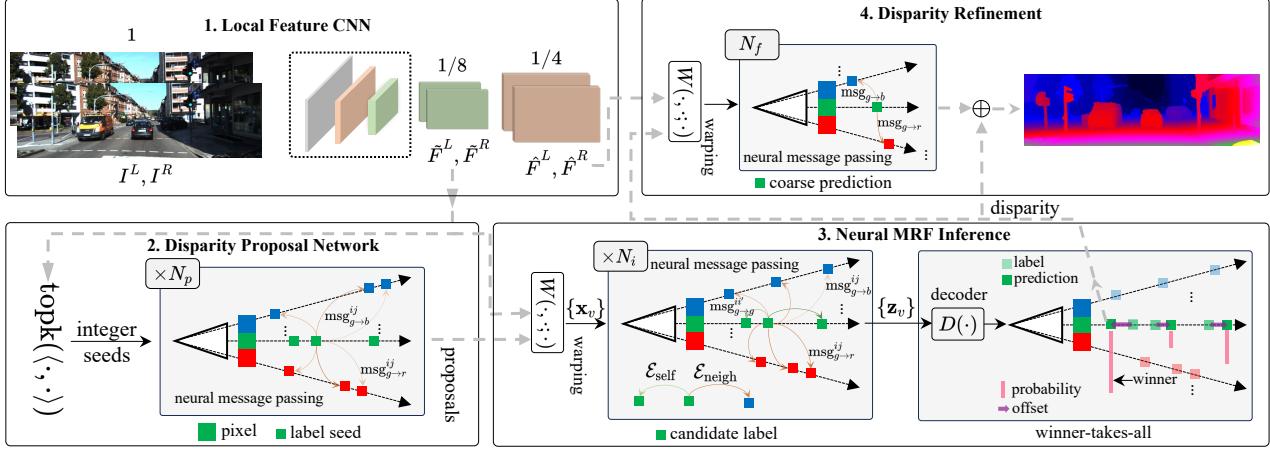


Figure 3. Overview of the proposed method. It has four components: **1.** A local feature CNN extracts the coarse and fine-level feature maps from the input image pair. **2.** A disparity proposal network prunes space of disparity. For every pixel, the top k disparity modals are identified, and then updated using N_p neural message passing, resulting in a sparse label set L_o . **3.** The MRF factorizes into a probabilistic graph, where each node corresponds to a candidate label and each edge connects a label pair from neighbor pixels. Different potential functions are used for intra- and inter-pixel label pairs respectively. The inferred latent embedding \mathbf{z}_v is then decoded to posterior probability and offset. The winner label is selected as the coarse prediction. **4.** Disparity refinement also leverages a neural MRF model but with only one label per pixel for efficiency. The inferred latent embeddings are decoded into disparity residuals.

learn these potential functions by inferring latent embeddings $\{\mathbf{z}_v\}$ that can explain the ground truth disparity map. **Observed label feature.** The observed feature of a candidate label must integrate matching cues from both left and right views. Given the coarse level features \tilde{F}^L and \tilde{F}^R , we compute the observed feature \mathbf{x}_v of a candidate label positioned at $\mathbf{p}_v := (i, j, z)$ using a warping function w.r.t. $\tilde{F}^L(i, j)$ and $\tilde{F}^R(i - z, j)$. We leave the formal definition of the warping function in the supplementary material.

3.2. Neural MRF Inference

Preliminaries: Embedding mean-field inference [12, 18]. Exact computation of posterior $p(\{\mathbf{z}_v\} | \{\mathbf{x}_v\})$ is intractable, even if Φ and Ψ are known and well-defined. The mean-field theory assumes that the posterior over latent variables can be factorized into \mathcal{V} independent marginals $q_v(\mathbf{z}_v)$, i.e., $p(\{\mathbf{z}_v\} | \{\mathbf{x}_v\}) \approx \prod_v q_v(\mathbf{z}_v)$. The optimal marginals $\{q_v\}$ under the mean-field assumption are obtained by minimizing the Kullback-Leibler (KL) divergence between the approximation posterior and the true posterior. Hamilton [18] shows that the optimal solution needs to satisfy the following fixed point equation for all $v \in \mathcal{V}$:

$$\log(q_v(\mathbf{z}_v)) = c_v + \log(\Phi(\mathbf{x}_v, \mathbf{z}_v)) + \sum_{u \in \mathcal{N}(v)} \int_{\mathbb{R}^d} q_u(\mathbf{z}_u) \log(\Psi(\mathbf{z}_u, \mathbf{z}_v)) d\mathbf{z}_u, \quad (3)$$

where c_v is a normalization constant and $\mathcal{N}(v)$ denotes the neighbor set of node v . It implies that marginal distribution $q_v(\mathbf{z}_v)$ is a function of node feature \mathbf{x}_v and neighbor marginals $q_u(\mathbf{z}_u), \forall u \in \mathcal{N}(v)$, i.e., $q_v(\mathbf{z}_v) = f(\mathbf{z}_v, \mathbf{x}_v, \{q_u\}_{u \in \mathcal{N}(v)})$. Supposing an injective feature map

ϕ , each marginal $q_v(\mathbf{z}_v)$ maps to an embedding $\mu_v = \int_{\mathbb{R}^d} q_v(\mathbf{z}_v) \phi(\mathbf{z}_v) d\mathbf{z}_v \in \mathbb{R}^d$. The solution of fixed point Eq. (3) in the embedding space could be approximated by iteratively evaluating

$$\mu_v^t = \tilde{f}(\mathbf{x}_v, \mu_v^{t-1}, \{\mu_u^{t-1}\}_{u \in \mathcal{N}(v)}), \quad (4)$$

where \tilde{f} is a vector-valued function and has complicated nonlinear dependencies on the potential functions Φ, Ψ , as well as the feature map ϕ .

Inference with neural message passing. Currently, mean-field inference is only tractable for restrictive potential functions, e.g., those from the exponential family. In our setting, it's hard to work out \tilde{f} in Eq. (4) since potential functions Φ, Ψ , and the feature map ϕ are unknown and need to be learned from data. To address this dilemma, we parameterize \tilde{f} to be a neural network that retains the inductive bias of mean-field inference. Notice that \tilde{f} exactly corresponds to some message passing operation as it *aggregates* information from neighbor embeddings (i.e., $\{\mu_u\}_{u \in \mathcal{N}(v)}$) and *updates* the node's current representation (i.e., μ_v^{t-1}).

The initial representation ${}^{(0)}\mu_v$ for label v is initialized with \mathbf{x}_v . Let ${}^{(\ell)}\mu_v$ be the intermediate embedding for label v at layer ℓ . The message $\mathbf{m}_{\mathcal{E}^\ell \rightarrow v}$ is the outcome of embedding aggregation along edges \mathcal{E}^ℓ , where $\mathcal{E}^\ell \in \{\mathcal{E}_{\text{self}}, \mathcal{E}_{\text{neigh}}\}$. The message passing update is formally defined as:

$$\begin{aligned} {}^{(\ell+1)}\hat{\mu}_v &= {}^{(\ell)}\mu_v + \mathbf{m}_{\mathcal{E}^\ell \rightarrow v} \\ {}^{(\ell+1)}\mu_v &= \text{MLP}\left({}^{(\ell+1)}\hat{\mu}_v\right) + {}^{(\ell+1)}\hat{\mu}_v, \end{aligned} \quad (5)$$

where the MLP conceptually corresponds to the unary potential function Φ since it controls the embedding update,

as shown in Eq. (3). Furthermore, we observe that potential function Ψ controls the weight to aggregate information from neighbor embeddings. Intuitively, pairwise potential $\Psi(\cdot, \cdot)$ measures *affinity* between connected labels. This observation, together with the success of Transformer, motivates us to leverage self-attention for the embedding aggregation. A fixed number of N_i layers are chained and alternatively aggregate along the neighbor and self edges.

Attentional aggregation. To aggregate embedding along the neighbor edges $\mathcal{E}_{\text{neigh}}$, a label v first extracts relevant information from all neighbors $\{u : (u, v) \in \mathcal{E}_{\text{neigh}}\}$, and then sums them up weighted by the affinity score. The message computation could be formulated as:

$$\begin{aligned} \mathbf{m}_{\mathcal{E}_{\text{neigh}} \rightarrow v} &= \sum_{(u, v) \in \mathcal{E}_{\text{neigh}}} \alpha_{uv} (\mathbf{v}_u + \mathbf{r}_{u-v}^v) \\ \alpha_{uv} &= \text{softmax}_u (\mathbf{q}_v^\top \mathbf{k}_u + \underbrace{\mathbf{q}_v^\top \mathbf{r}_{u-v}^k + \mathbf{k}_u^\top \mathbf{r}_{u-v}^q}_{\text{content-adaptive positional bias}}), \end{aligned} \quad (6)$$

where $\mathbf{q}_v, \mathbf{k}_v, \mathbf{v}_v$ are the *query*, *key* and *value* get by a linear projection of the embedding $(\ell)\mu_v$, and $\mathbf{r}_{u-v}^q, \mathbf{r}_{u-v}^k, \mathbf{r}_{u-v}^v$ are encodings of the relative position $\mathbf{p}_u - \mathbf{p}_v$ in three different subspaces. Adding relative positional encoding \mathbf{r}_{u-v}^v to the value vector, messages will become position-dependent. This will benefit disparity aggregation in ambiguous areas, where the relative position is important. The query \mathbf{q}_v and key \mathbf{k}_u could contain information about where to focus in the neighborhood. Thus, their dot product with the relative positional encodings are utilized as content-adaptive positional bias of self-attention. STTR [28] has found similar positional bias is beneficial for stereo matching.

The 3D relative position $\mathbf{p}_u - \mathbf{p}_v$ consists of pixel coordinate difference $(\Delta i, \Delta j)$, and disparity difference Δz . In our setting, however, it is memory unaffordable to directly encode $\mathbf{p}_u - \mathbf{p}_v$ in the sinusoidal format or through a small network [63] due to the quadratic pairwise combination. Thus, we pursue an approximate encoding method. Given that label v and u come from a pair of pixels within a $M \times M$ window, the pixel coordinate difference $(\Delta i, \Delta j)$ will take integer values in the range $[-M + 1, M - 1]$ along each axis. With this in mind, we retrieve the position encoding of $(\Delta i, \Delta j)$ from a learnable embedding table $\mathbf{P} \in \mathbb{R}^{3 \times (2M-1) \times (2M-1) \times D}$, where the leading dimension 3 is for query, key and value respectively, and D is the number of channels. For the encoding of relative disparity Δz , we approximate with sinusoidal encoding of absolute disparity. Specifically, we concatenate it with embedding $(\ell)\mu_v$ before the linear projection to get query, key and value.

For embedding aggregation along self edges $\mathcal{E}_{\text{self}}$, the pixel coordinate difference $(\Delta i, \Delta j)$ are always zero, and thus we omit the related position encoding terms in Eq. (6).

Disparity estimation. After N_i layers' neural message passing, we estimate disparity map by decoding the inferred

latent embedding $\{(N_i)\mu_v\}$. For every pixel on the coarse level, the latent embedding of each candidate label is decoded into 8×8 disparity offsets and 8×8 probabilities. The 8×8 disparity offsets w.r.t. the candidate label are used to compute disparity hypotheses for the 8×8 pixels in the original image. As a result, we produce k , the number of candidate labels, scored (with probability) disparity hypotheses for every pixel in the input image. We estimate the disparity using the winner-takes-all strategy, as depicted in Fig. 3.

3.3. Disparity Proposal Network

To ensure tractable neural MRF inference, we propose a Disparity Proposal Network (DPN) to provide each pixel o with a *small* candidate label space L_o . The DPN first identifies the top k disparity modals in the range $[0, z_{\max}]$. Then, it updates them by leveraging the inherent spatial coherence, resulting in k candidate labels for each pixel.

Top k label seeds. At pixel $o = (i, j)$, the initial matching score for integer disparity $z \in [0, z_{\max}]$ is computed using the inner product $\langle \tilde{F}^L(i, j), \tilde{F}^R(i, j - z) \rangle$. We identify the disparity modals, *i.e.*, the integer disparities with locally maximum matching scores, using a 1D *max pooling* along the disparity dimension. The kernel size is set to 3 so as to detect the local maximum within the vicinity of $[-1, 1]$. Then a topk operation identifies the best k modals. Each modal is characterized by its *position* \mathbf{p} and *matching feature* \mathbf{d} (details in supplementary). We refer to them jointly (\mathbf{p}, \mathbf{d}) as the *label seed*. The position consists of i and j coordinates as well as the disparity z , *i.e.*, $\mathbf{p}_v := (i, j, z)$.

Label seeds propagation. The top k label seeds may fail to capture the true disparities, particularly in the textureless or occluded regions. Even so, good label seeds still make up the majority in the entire image field. Our intuition is to rectify erroneous label seeds using the information from dependent good label seeds. To facilitate this, we also utilize a message passing network to exchange matching features between label seeds. It is well-known that long-range dependency is critical for tackling occlusion and large textureless regions. To efficiently exchange matching features with distant label seeds, we implement message aggregation using a cross-shaped window self-attention [13]. In contrast to neural MRF inference, DPN ignores *self* edges since intra-pixel competition does not make sense in the proposal extraction stage. The enhanced matching features of each label seed is decoded into residuals w.r.t. the initial integer disparity. More details about the cross-shaped attentional message aggregation are provided in the supplementary material.

3.4. Disparity Refinement

Thus far, the neural MRF inference on the coarse level features already exhibits competitive performance (Tab. 5). It can be further improved by performing N_f neural message passing on the fine level features for detailed structure refinement. For efficiency, we use only one candidate label for

Methods	PSMNet [6]	GANet-deep [61]	AANet [56]	LEAStereo [11]	ACVNet [54]	IGEV-Stereo [55]	Ours
EPE [px]	1.09	0.78	0.87	0.78	0.48	0.47	0.45
Bad 1.0 [%]	12.1	8.7	9.3	7.82	5.02	5.35	4.50

Table 1. Quantitative evaluation on SceneFlow test set. Our method achieves state-of-the-art performance on both metrics. **Bold**: Best.

every pixel on the fine level. The candidate label is obtained by reducing the coarse disparity estimation using a strided-4 median pooling with kernel size 4×4 . Because of a single candidate label, the refinement differs from coarse inference (Sec. 3.2) in two aspects: 1) the graph is free of *self* edges, 2) posterior probability branch is no longer needed.

3.5. Loss Functions

To generate ground truth for proposal extraction at $1/8$ resolution, we propose a superpixel-guided disparity map down-sample operator, which reduces each non-overlapping 8×8 patch to multiple modals (details in supplementary).

Proposal loss. We expect candidate labels will identify all ground truth disparity modals $\{z_k^*\}$. To measure this, our loss finds an optimal bipartite matching between candidate labels and ground truth modals, and then optimizes the one-to-one matching. The candidate label set of pixel o , denoted by $L_o = \{\hat{z}_k\}$, is associated with proposal loss as:

$$L^{\text{prop}} = \sum_k \text{Smooth}_{L1}(z_k^* - \hat{z}_{\hat{o}(k)}) \quad (7)$$

where $\hat{o}(k)$ is the index of candidate labels that has been matched with z_k^* . This loss is inspired by DETR [4], a pioneering work in Transformer-based object detection.

Disparity loss. Given the ground truth disparity z^* , the disparity loss is formulated as:

$$L^{\text{disp}} = \sum_{z'} p(z') |z' - z^*|, \quad (8)$$

where $\{z'\}$ and $\{p(z')\}$ are the disparity hypotheses and posterior probabilities inferred by neural MRF.

4. Experiments

We evaluate the proposed stereo models on two widely used datasets, SceneFlow [33] and KITTI [15, 34]. SceneFlow is a synthetic dataset which provides 35,454 training and 4,370 testing pairs of size 960×540 with accurate ground truth disparity maps. KITTI 2012 and 2015 are real-world datasets with sparse LIDAR ground truth disparities for the training set. KITTI 2012 includes 194 training and 195 testing pairs, while KITTI 2015 has 200 training and 200 testing pairs. Additionally, we use the training pairs of KITTI, Middlebury 2014 [38] and ETH3D [39] to evaluate zero-shot generalization ability.

4.1. Implementation Details

We implement our model using the PyTorch framework on NVIDIA RTX 3090 GPUs. The AdamW optimizer is used

in conjunction with a one-cycle learning rate scheduler for all training. On the SceneFlow dataset, we train models on 384×768 random crops for 300k steps with a batch size of 8 and set the maximum learning rate to 0.0005. Following the standard protocol, we exclude all pixels with ground truth disparity greater than 192 from the evaluation. For our submissions to the KITTI benchmark, the model pre-trained on SceneFlow is fine-tuned on the mixed KITTI 2012 and 2015 training sets for another 39k steps with a batch size of 4. We randomly crop images to 304×1152 and use a maximum learning rate of 0.0002 for fine-tuning. More implementation details are provided in the supplementary material.

4.2. Benchmark Evaluation

We compare our stereo models with the published state-of-the-art methods on SceneFlow “finalpass”, KITTI 2012 and KITTI 2015 datasets. The evaluation metrics on the SceneFlow test set are average end point errors (EPE) and bad pixel ratio with 1 pixel threshold (Bad 1.0). Table 1 illustrates that our model achieves state-of-the-art performance on both metrics. Our method outperforms LEAStereo [11] and ACVNet [54] by 42.5% and 10.4% on the outlier ratio under 1 pixel threshold (Bad 1.0), respectively. These two are current state-of-the-art local methods [37] that regularize cost volume using 3D convolution networks.

We also evaluate our method on the test set of KITTI 2012 and 2015, and the results are submitted to the [online leaderboard](#). At the time of writing, our method ranks 1st on both datasets among all published methods while running faster than 100 ms. As shown in Tab. 2, we achieve the best performance for almost all metrics on KITTI 2012 and 2015. Compared to prior leading global stereo methods, *e.g.*, PBCP [40] and LBPS [25], our method significantly outperforms them by more than 50%. We present some qualitative results in Fig. 4. Our method performs well in large textureless and detailed regions. In addition, we compare the point clouds generated by our approach and the top-performing LEAStereo [11]. As depicted in Fig. 2a, our approach can greatly alleviate *bleeding artifacts* [51] and produce sharp disparity estimates near discontinuities.

4.3. Zero-shot Generation

Generalizing from synthetic training data to unseen real-world datasets is crucial because collecting large-scale real-world datasets for training is challenging and expensive. In this evaluation, we use the model trained only on synthetic SceneFlow dataset, *i.e.*, that reports the accuracy of Tab. 1, and directly test it on the KITTI, Middlebury and ETH3D training sets. Table 3 compares our approach with current

Methods	KITTI 2012					KITTI 2015			
	bad 2.0		bad 3.0		EPE	BG	FG	ALL	Runtime
	noc [%]	all [%]	noc [%]	all [%]	noc [px]	All Areas [%]		[s]	
GCNet [22]	2.71	3.46	1.77	2.30	0.6	2.21	6.16	2.87	0.9
PSMNet [6]	2.44	3.01	1.49	1.89	0.5	1.86	4.62	2.32	0.41
GwcNet [17]	2.16	2.71	1.32	1.70	0.5	1.74	3.93	2.11	0.32
GANet-deep [61]	1.89	2.50	1.19	1.60	0.4	1.48	3.46	1.81	1.8
CSPN [10]	1.79	2.27	1.19	1.53	-	1.51	<u>2.88</u>	1.74	1.0
RAFT-Stereo [30]	1.92	2.42	1.30	1.66	0.4	1.58	3.05	1.82	0.38
LEAStereo [11]	1.90	2.39	1.13	1.45	0.5	1.40	2.91	1.65	0.3
ACVNet [54]	1.83	2.35	1.13	1.47	0.4	<u>1.37</u>	3.07	1.65	0.2
IGEV-Stereo [55]	1.71	<u>2.17</u>	1.12	1.44	0.4	1.38	2.67	1.59	0.18
PCWNet [43]	1.69	2.18	<u>1.04</u>	<u>1.37</u>	0.4	<u>1.37</u>	3.16	1.67	0.44
PBCP [40]	3.63	5.01	2.36	3.45	0.7	2.58	8.74	3.61	68
LBPS [25]	-	-	-	-	-	2.85	6.35	3.44	0.39
Ours	1.59	2.07	1.01	1.35	0.4	1.28	3.13	1.59	0.09

Table 2. **Quantitative evaluation on KITTI 2012 and 2015.** For KITTI 2012, we report the outlier ratio with error greater than x pixels (bad x) in both non-occluded (noc) and all regions (all), as well as the overall EPE in non-occluded pixels. For KITTI 2015, we report D1 metric in background regions (BG), foreground regions (FG), and all. **Bold:** Best, Underline: Second best.

Methods	KITTI-12	KITTI-15	Middlebury	ETH3D
DSMNet [62]	6.2	6.5	<u>8.1</u>	6.2
RAFT-Stereo [30]	4.7	5.5	9.4	3.3
CREStereo++ [21]	4.7	<u>5.2</u>	-	4.4
IGEV-Stereo [55]	5.2	5.7	8.8	4.0
Ours	4.2	5.1	7.5	<u>3.8</u>

Table 3. **Zero-shot generalization evaluation.** All methods are only trained on SceneFlow and evaluated based on the outlier ratio with error greater than the specific threshold. We use the standard thresholds: D1 for KITTI, 2px for Middlebury, 1px for ETH3D.

robust methods². It turns out that our approach even surpasses DSMNet [62] and CREStereo++ [21], both of which are specifically designed for cross-domain generalization. Fig. 5 shows some qualitative results for zero-shot generation on ETH3D [39] and Middlebury [38].

4.4. Ablation Studies

Effectiveness of disparity proposal network. The upper limit of our architecture is determined by the quality of candidate labels. In this study, we evaluate the recall and accuracy of candidate labels. Two metrics are used: x -pixels recall, the percent of pixels with ground truth identified by candidate labels under the threshold x pixels; and the average end point errors (EPE) between the best label and ground truth, as visualized in Fig. 2b. Tab. 4 indicates that the DPN attains a recall of 99.8% on popular datasets with 8-pixel threshold, which is equivalent to 1 pixel at the coarse level features. Moreover, the EPE metric of candidate labels is substantially lower than that of final estimation, e.g., 0.22

²IGEV-Stereo [55] uses a non-standard evaluation protocol. This paper reevaluates it using the standard protocol to compare with other methods.

Datasets	candidate labels			label seeds	
	3px [%]↑	8px [%]↑	EPE [px]↓	8px [%]↑	16px [%]↑
SceneFlow-test	99.29	99.77	0.22	91.94	94.85
KITTI-12 val	98.88	99.72	0.31	93.87	95.89
KITTI-15 val	99.66	99.95	0.21	93.91	95.92

Table 4. **Quantitative evaluation of disparity proposal.** The validation set of KITTI 2012 and 2015 both contain 40 image pairs. We report EPE and the x -pixel recall rate. The extracted candidate labels have significantly better quality than initialized label seeds.

vs. 0.45 in SceneFlow. It implies that the DPN is not the performance bottleneck of our proposed architecture.

Number of candidate labels. The results in Tabs. 1 to 4 are all achieved with the number of candidate labels $k = 4$. We investigate the impact of the number of candidate labels on model accuracy, generalization ability, and inference time vary with. As shown in Fig. 6a, the proposed architecture is not sensitive to the number of candidate labels. Our method even achieves competitive performance with only one candidate label. Furthermore, our neural MRF is robust to occasional DPN failures (Fig. 2b). More candidate labels do not always result in better performance, since it increases the risk of choosing an incorrect hypothesis. As Fig. 6a suggests, $k = 2$ may be a good choice for real time stereo matching with marginal sacrifice of performance.

Self edges. Message passing along self edges is usually overlooked in hand-crafted stereo MRF and convolutional cost aggregation. This paper uncovered its significance in neural MRF inference. As shown in Fig. 6b, direct information exchange along self edges makes the winner label more prominent, such that the loss in Eq. (8) is more focused on

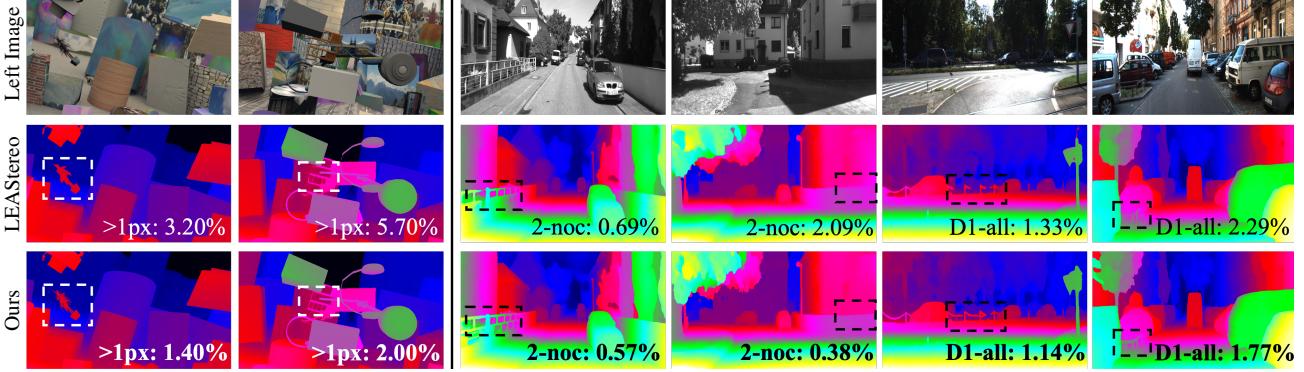


Figure 4. **Qualitative results on SceneFlow [33] and KITTI [15, 34] benchmarks.** The leftmost two columns show results on SceneFlow, while the middle two and the rightmost two columns show results on KITTI 2012 and KITTI 2015, respectively. Our method exhibits outstanding performance in large textureless and detailed regions, compared with the top-performing LEAStereo [11].

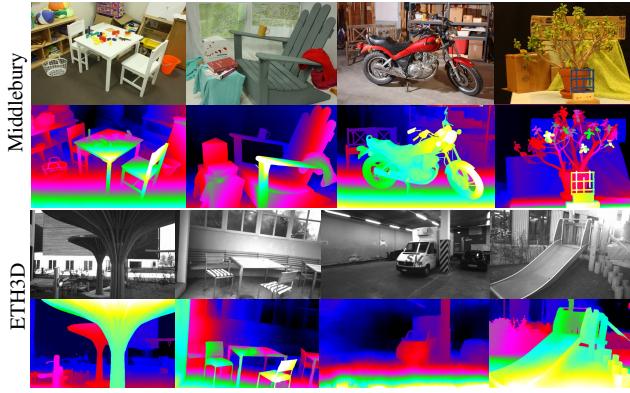


Figure 5. Zero-shot generalization on ETH3D and Middlebury.

	EPE[px]	Bad1.0[%]	Time[s]
baseline	0.58	5.96	0.064
w/ adaptive bias	0.57	5.80	0.067
w/ position aggregation	0.57	5.53	0.072
w/ <i>shared</i> self edge	0.56	5.48	0.072
w/ self edge	0.53	5.24	0.081
w/ refinement	0.45	4.50	0.101
w/ refinement × 2	0.43	4.13	0.121

Table 5. Ablation studies to investigate the effect of individual components on SceneFlow test. The baseline uses a fixed 2D relative positional bias [31]. The 2nd last row is our full model.

the disparity estimation of the winner label. This brings considerable improvements, *e.g.*, lowering the EPE metric by 7% (0.53 vs. 0.57), detailed in Tab. 5. We employ separate potential functions for self and neighbor edges. We validate this design by comparing it with the variant that uses the same function, denoted as *shared* self edges in Tab. 5. Our design consistently outperforms the alternative.

Attentional aggregation components. We ablate adaptive bias and position aggregation in Tab. 5. They both considerably reduce outliers. Adaptive positional bias performs better for edge pixels. We conjecture it brings benefits to the sharp depth boundaries shown in Fig. 2a. By injecting

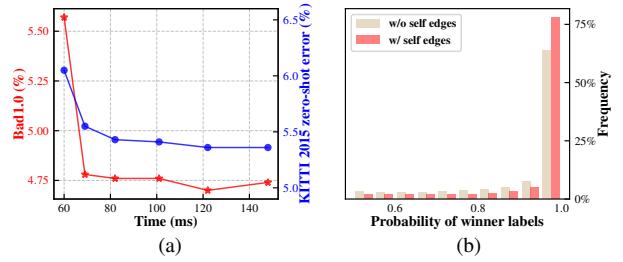


Figure 6. (a) Tradeoff between time, accuracy, and generalization. From left to right, the number of candidate labels are 1,2,3,4,5,6, respectively. (b) Histogram of probabilities of winner labels. Message passing along *self* edges makes the winner more prominent. The statistics are based on 4.5M uniformly sampled pixels.

positional encoding to *value*, position aggregation results in superior estimation in large textureless regions (see Fig. 4).

Refinement. Among all components in Tab. 5, refinement contributes the most, as the fine level ($1/4$) features would offer valuable information for detailed structures and pixels near boundaries. If we cascade two refinement modules together, better performance can be achieved with the compromise of generalization ability. This enables users to balance performance and generalization based on their needs.

5. Conclusion

We proposed a novel Neural MRF (NMRF) formulation for stereo matching and demonstrated its strong performance and generalization ability. It is distinct from the learning architectures currently in use, *e.g.*, convolutional cost aggregation and iterative recurrent refinement. We hope our new perspective will provide a new paradigm for stereo matching and can be extended to similar tasks, *e.g.*, optical flow.

Acknowledgement. This work is supported by Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, the RGC grant (14207320) from Hong Kong SAR government, CUHK T-Stone Robotics Institute, and the Hong Kong Center for Logistics Robotics.

References

- [1] Stan Birchfield and Carlo Tomasi. Multiway cut for stereo and motion with slanted surfaces. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 489–495. IEEE, 1999. [1](#)
- [2] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, pages 1–11, 2011. [2](#)
- [3] Guillermo Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6247–6257, 2020. [2](#)
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. [6](#)
- [5] Rohan Chabra, Julian Straub, Christopher Sweeney, Richard Newcombe, and Henry Fuchs. Stereodrnet: Dilated residual stereonet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11786–11795, 2019. [2](#)
- [6] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. [2, 6, 7](#)
- [7] Chuangrong Chen, Xiaozhi Chen, and Hui Cheng. On the over-smoothing problem of cnn based disparity estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8997–9005, 2019. [3](#)
- [8] Zhuoyuan Chen, Xun Sun, Liang Wang, Yinan Yu, and Chang Huang. A deep visual correspondence embedding model for stereo matching costs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 972–980, 2015. [1, 2](#)
- [9] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. [3](#)
- [10] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 2019. [2, 7](#)
- [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. *Advances in Neural Information Processing Systems*, 33, 2020. [1, 2, 3, 6, 7, 8](#)
- [12] Hanjun Dai, Bo Dai, and Le Song. Discriminative embeddings of latent variable models for structured data. In *International conference on machine learning*, pages 2702–2711. PMLR, 2016. [2, 4](#)
- [13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. [3, 5, 2](#)
- [14] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4384–4393, 2019. [3](#)
- [15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. [2, 6, 8](#)
- [16] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017. [2](#)
- [17] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. [2, 7](#)
- [18] William L Hamilton. *Graph representation learning*. Morgan & Claypool Publishers, 2020. [2, 4](#)
- [19] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3279–3286, 2015. [2](#)
- [20] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. [1, 2](#)
- [21] Junpeng Jing, Jiankun Li, Pengfei Xiong, Jiangyu Liu, Shuaicheng Liu, Yichen Guo, Xin Deng, Mai Xu, Lai Jiang, and Leonid Sigal. Uncertainty guided adaptive warping for robust and efficient stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3318–3327, 2023. [1, 2, 7](#)
- [22] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. [1, 2, 7](#)
- [23] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 573–590, 2018. [2](#)
- [24] Patrick Knobelreiter, Christian Reinbacher, Alexander Shekhovtsov, and Thomas Pock. End-to-end training of hybrid cnn-crf models for stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2339–2348, 2017. [1, 2](#)
- [25] Patrick Knobelreiter, Christian Sormann, Alexander Shekhovtsov, Friedrich Fraundorfer, and Thomas Pock. Belief propagation reloaded: Learning bp-layers for labeling problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7900–7909, 2020. [1, 2, 6, 7](#)
- [26] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions*

- on Pattern Analysis and Machine Intelligence*, 44(4):1738–1764, 2020. 2
- [27] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022. 1, 2
- [28] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unterath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021. 5
- [29] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2811–2820, 2018. 2
- [30] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 1, 2, 7
- [31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 8
- [32] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016. 1, 2
- [33] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1, 2, 6, 8
- [34] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 2, 6, 8
- [35] Moritz Menze, Christian Heipke, and Andreas Geiger. Discrete optimization for optical flow. In *Pattern Recognition: 37th German Conference, GCPR 2015, Aachen, Germany, October 7–10, 2015, Proceedings 37*, pages 16–28. Springer, 2015. 3
- [36] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2
- [37] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. 1, 6
- [38] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2–5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 2, 6, 7
- [39] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 2, 6, 7
- [40] Akihito Seki and Marc Pollefeys. Patch based confidence prediction for dense disparity map. In *BMVC*, page 4, 2016. 2, 6, 7
- [41] Akihito Seki and Marc Pollefeys. Sgm-nets: Semi-global matching with neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 231–240, 2017. 2
- [42] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, 2021. 1, 2, 3
- [43] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *European Conference on Computer Vision*, pages 280–297. Springer, 2022. 1, 2, 7
- [44] Ron Slossberg, Aaron Wetzler, and Ron Kimmel. Deep stereo matching with dense crf priors. *arXiv preprint arXiv:1612.01725*, 2016. 1
- [45] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2
- [46] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE transactions on pattern analysis and machine intelligence*, 30(6):1068–1080, 2008. 1, 3
- [47] Tatsunori Taniai, Yasuyuki Matsushita, Yoichi Sato, and Takeshi Naemura. Continuous 3D Label Stereo Matching using Local Expansion Moves. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(11):2725–2739, 2018. 1, 2
- [48] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021. 2
- [49] Michael Tanner, Pedro Pinies, Lina Maria Paz, Ştefan Săftescu, Alex Bewley, Emil Jonasson, and Paul Newman. Large-scale outdoor scene reconstruction and correction with vision. *The International Journal of Robotics Research*, 41(6):637–663, 2022. 1

- [50] Tappen. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 900–906. IEEE, 2003. 1
- [51] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8942–8952, 2021. 6
- [52] Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. *Advances in neural information processing systems*, 31, 2018. 2
- [53] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1
- [54] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12981–12990, 2022. 2, 6, 7
- [55] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 1, 2, 6, 7
- [56] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 2, 6
- [57] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2015. 2
- [58] Jure Zbontar, Yann LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318, 2016. 1, 2
- [59] Nadia Zenati and Noureddine Zerhouni. Dense stereo matching with application to augmented reality. In *2007 IEEE International Conference on Signal Processing and Communications*, pages 1503–1506. IEEE, 2007. 1
- [60] Chi Zhang, Zhiwei Li, Yanhua Cheng, Rui Cai, Hongyang Chao, and Yong Rui. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2057–2065, 2015. 1
- [61] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 1, 2, 6, 7
- [62] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Europe Conference on Computer Vision (ECCV)*, 2020. 1, 2, 7
- [63] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 5

Neural Markov Random Field for Stereo Matching

Supplementary Material

1. Implementation Details of Loss Functions

Superpixel-guided disparity downsample. To provide supervision signal for disparity proposal extraction at $1/8$ resolution, we introduce a superpixel-guided disparity map downsample function, which reduces each 8×8 disparity window to multiple modals. We divide the ground truth disparity map into non-overlapping 8×8 windows and perform an independent downsample for each window.

First, we over-segment the left image I^L into superpixels using the LSC method implemented in OpenCV³. As shown in Fig. 7, the superpixel effectively groups adjacent pixels while preserving local image structures, making it appropriate for reducing disparity values. Subsequently, each 8×8 window is decomposed into multiple segments utilizing the superpixel label map. We sort the segments based on their pixel count and compute the *median* disparity of each segment as the representative. To mitigate over-segmentation in the window, we employ a non-maximum suppression (NMS) on the representative disparity list. The suppression criterion is based on the difference between representative disparity values. If the absolute difference is less than 0.5 pixels, we merge the suppressed segment into the segment that suppresses it. After the merge step, we sort the segments again based on the pixel count and choose the median disparity of the top 4 segments as the downsample function output. If there exist fewer than 4 segments, we pad the output with null values.

Proposal loss. Once we have obtained the downsampled ground truth disparity modals, we use it to train our disparity proposal extraction network, as detailed in Sec. 3.5 of the paper. When computing the proposal loss in Eq. (7), we need to find the optimal bipartite matching between proposals and ground truth modals. For instance, consider a pixel on the coarse level with four ground truth disparity modals, namely $\{1.1, 1.8, \phi, \phi\}$, and four extracted proposals, namely $\{1.4, 10.2, 10.8, 11.2\}$. Ignoring null value ϕ , the optimal bipartite matching pairs consist of $(1.1, 1.4)$ and $(1.8, 10.2)$. However, we need to be careful with the close ground truth modals. In this case, the proposal 1.4 already captures the two close ground truth modals 1.1 and 1.8. Thus, the matching pair $(1.8, 10.2)$ is unnecessary and may induce negative impact on the training.

To address this, we perform an online non-maximum suppression (NMS) on ground truth modals when computing the proposal loss. First, we sort ground truth modals based on their proximity to the extracted proposal set. Proximity is measured by the minimum distance between the

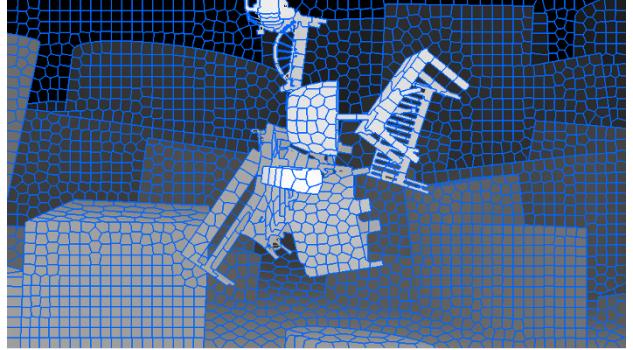


Figure 7. Superpixels overlaid on ground truth disparity map.

ground truth modal and all proposals. Then, we suppress the close ground truth modals using the threshold of 8 pixels. In continuation with the above example, our online NMS reduces the ground truth modals to $\{1.1, \phi, \phi, \phi\}$, and only one matching pair $(1.1, 1.4)$ is leveraged for proposal loss.

Initialization loss. Besides the proposal loss, an initialization loss is also employed to supervise label seeds to identify ground truth modals. As described in Sec. 3.3, label seeds are derived from a 3D cost volume \mathbf{C} , with

$$\mathbf{C}(i, j, z) = \langle \tilde{F}^L(i, j), \tilde{F}^R(i - z, j) \rangle. \quad (9)$$

We expect the initialization loss to penalize the discrepancy between ground truth modals and the 3D cost volume \mathbf{C} . To this end, we transform the ground truth modals of each pixel o into a probability distribution, $p^*(z) = \sum_k w_k \delta(z - z_k^*)$, where $\{z_k^*\}$ are the ground truth modals at pixel o and δ is the Dirac delta function. The mass weights $\{w_k\}$ are empirically set to $\{0.5, 0.3, 0.1, 0.1\}$ for the four sorted ground truth modals. This simple strategy performs well in all our experiments. Note that ground truth modals are given with subpixel precision, however, label seed extractioin happens with integer disparities. Therefore, we displace the probability mass as z_k^* to nearby integer disparities as

$$\begin{aligned} \tilde{p}^*(z) = & \sum_k w_k (\lfloor z_k^* \rfloor + 1 - z_k^*) \delta(z - \lfloor z_k^* \rfloor) + \\ & w_k (z_k^* - \lfloor z_k^* \rfloor) \delta(z - \lfloor z_k^* \rfloor - 1). \end{aligned} \quad (10)$$

We define the initialization loss to be the cross entropy between ground truth probability \tilde{p}^* and softmax of 3D cost volume \mathbf{C} along the z dimension, *i.e.*,

$$L^{\text{init}} = - \sum_{z \in [0, z_{\max}]} \tilde{p}^*(z) \cdot \log(\text{softmax}_z(\mathbf{C}(i, j, z))). \quad (11)$$

³<https://opencv.org>

2. Additional Implementation Details

Local feature CNN. We use a similar backbone as RAFT-Stereo [30], which consists of a strided-2 stem and three residual blocks with strides 1, 2, 1, respectively. The network produces a feature map with 128 channels at $1/4$ input image resolution, which is then downsampled through average pooling with a stride of 2 and a kernel size of 2. We further pass the obtained $1/8$ resolution feature map and the original $1/4$ resolution feature map to a shared convolution layer with 256 channels.

Neural message passing. The number of message passing blocks we use in label seeds propagation (N_p), MRF inference (N_i), and refinement (N_f) are 5, 10, 5 respectively. We use same settings for all experiments. The channels of embedding vectors in all message passing blocks are always 128. The neighborhood window size is 4×4 for refinement (Sec. 3.4), and 6×6 for neural MRF inference (Sec. 3.2). We also found that more message passing blocks and larger window size would bring slightly better accuracy with considerable computation overhead.

Observed label feature. The observed feature of a candidate label must integrate matching cues from both left and right views. Given the coarse level features \tilde{F}^L and \tilde{F}^R , we compute the observed feature \mathbf{x}_v of a candidate label positioned at $\mathbf{p}_v := (i, j, z)$ using a warping function w.r.t. $\tilde{F}^L(i, j)$ and $\tilde{F}^R(i - z, j)$ as

$$\begin{aligned}\mathbf{x}_v &= \text{MLP}(\mathbf{x}_v^{\text{concat}} \parallel \mathbf{x}_v^{\text{corr}}) \\ \mathbf{x}_v^{\text{concat}} &= \gamma_1\left(\tilde{F}^L(i, j)\right) \parallel \gamma_1\left(\tilde{F}^R(i - z, j)\right) \\ \mathbf{x}_v^{\text{corr}} &= \frac{N_g}{N_c} \langle \gamma_2\left(\tilde{F}_g^L(i, j)\right), \gamma_2\left(\tilde{F}_g^R(i - z, j)\right) \rangle,\end{aligned}\quad (12)$$

where $[\cdot] \parallel [\cdot]$ denotes concatenation along the channel dimension. $\tilde{F}_g^L, \tilde{F}_g^R$ are g^{th} grouped features of F^L and F^R , which are evenly divided into N_g groups. N_c is the channel of coarse level features, and $\langle \cdot, \cdot \rangle$ denotes the inner product. γ_1 and γ_2 are normalization functions to make the terms $\mathbf{x}_v^{\text{concat}}$ and $\mathbf{x}_v^{\text{corr}}$ share similar data distribution. Both γ_1 and γ_2 consist of two linear layers, with instance normalization and activation function following the first linear layer. Since disparity z is a real number, we leverage bilinear interpolation when indexing feature map \tilde{F}^R . Our formulation is inspired by the success of GwcNet [17] and PCWNet [43].

In the refinement stage, we use the same warping function as Eq. (12), but w.r.t. fine level features \hat{F}^L and \hat{F}^R .

Cross-shaped window attention. To efficiently capture long-range dependency for label seed message exchange, we employ the cross-shaped attention mechanism proposed in CSWin Transformer [13]. As illustrated in Fig. 8a, the interested label seed, positioned at $\mathbf{p}_v := (i, j, z_k)$, aggregates matching information from all other label seeds that share the same i or j coordinate. We follow the parallel multi-head grouping strategy and locally-enhanced posi-

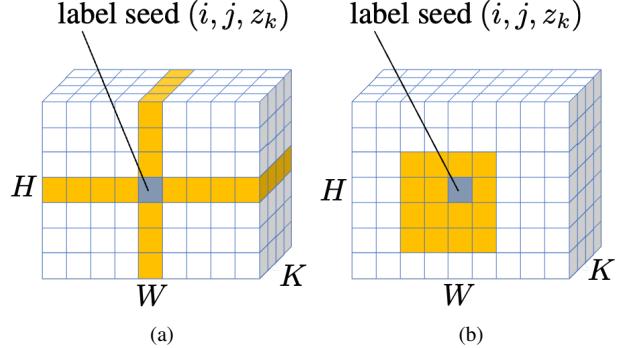


Figure 8. (a) Cross-shaped attention window arrangement, (b) local attention window arrangement.

	candidate labels		disparity estimation	
	3px [%]↑	8px [%]↑	EPE [px]↓	Bad 1.0 [%]↓
	cross-shaped	99.29	99.77	0.45
local window	99.18	99.72	0.47	4.56

Table 6. Performance comparison between cross-shaped window attention and local window attention in label seed propagation.

tional encoding of CSWin Attention [13] when performing attentional aggregation. The initial matching feature ${}^{(0)}\mathbf{d}_v$ of a label seed v is expected to encode cost features and underlying disparity value, formally defined as:

$${}^{(0)}\mathbf{d}_v = \text{MLP}\left(\gamma_3(L_z(\mathbf{C}(i, j, :))) \parallel \text{PE}(z)\right), \quad (13)$$

where \mathbf{C} denotes the 3D cost volume computed in label seeds extraction using Eq. (9). The lookup operator L_z retrieves cost features from volume slice $\mathbf{C}(i, j, :)$ around integer disparity z for pixel (i, j) , akin to RAFT-Stereo [30]. We apply a two-layer MLP called γ_3 to normalize the retrieved cost features before concatenating it with the sinusoidal positional encoding (PE) of disparity z .

We validate the design of cross-shaped window attention by comparing with the local window attention shown in Fig. 8b. The local window size is set to 8×8 to match the computation complexity of cross-shaped window attention on SceneFlow dataset [33]. The results are shown in Tab. 6. Due to its ability to capture long-range dependency, cross-shaped window attention performs better than local window attention for label seed propagation. However, we do not adopt the cross-shaped window attention in neural MRF inference and refinement, since it does not adapt to our proposed content-adaptive positional bias and position aggregation for different input resolutions.