H-Net: Unsupervised Attention-based Stereo Depth Estimation Leveraging Epipolar Geometry

Baoru Huang[1*] Jian-Qing Zheng[2,3*†] Stamatia Giannarou[1]

Daniel S. Elson[1]
[1]The Hamlyn Centre for Robotic Surgery, Imperial College London, U.K.
[2]The Kennedy Institute of Rheumatology, University of Oxford, U.K.
[3]Big Data Institute, University of Oxford, U.K.

Abstract

Depth estimation from a stereo image pair has become one of the most explored applications in computer vision, with most previous methods relying on fully supervised learning settings. However, due to the difficulty in acquiring accurate and scalable ground truth data, the training of fully supervised methods is challenging. As an alternative, self-supervised methods are becoming more popular to mitigate this challenge. In this paper, we introduce the H-Net, a deep-learning framework for unsupervised stereo depth estimation that leverages epipolar geometry to refine stereo matching. For the first time, a Siamese autoencoder architecture is used for depth estimation which allows mutual information between rectified stereo images to be extracted. To enforce the epipolar constraint, the mutual epipolar attention mechanism has been designed which gives more emphasis to correspondences of features that lie on the same epipolar line while learning mutual information between the input stereo pair. Stereo correspondences are further enhanced by incorporating semantic information to the proposed attention mechanism. More specifically, the optimal transport algorithm is used to suppress attention and eliminate outliers in areas not visible in both cameras. Extensive experiments on KITTI2015 and Cityscapes show that the proposed modules are able to improve the performance of the unsupervised stereo depth estimation methods while closing the gap with the fully supervised approaches.

*Baoru Huang and Jian-Qing Zheng contribute equally to this paper
†jianqing.zheng@kennedy.ox.ac.uk

1. Introduction

Humans are remarkably capable of inferring the 3D structure of a real world scene even over short timescales. For example, when navigating along a street, we are able to locate obstacles and vehicles in motion and avoid them with a fast response time. Years of substantial interest in geometric computer vision has not yet accomplished comparable modeling capabilities to humans for real-world scenes where reflections, occlusions, non-rigidity and textureless areas exist. So what can human ability be attributed to? A central concept is that humans learn the regularities of the world while interacting with it, moving around, and observing vast quantities of scenes. Consequently, we develop a rich, consistent and structural understanding of the world, which is utilized when we perceive a new scene. Our binocular vision is one supporting feature, from which the brain can not only build disparity maps, but can also combine to obtain structural information. These two ideas can help to solve one of the fundamental problems in computer vision — depth estimation — whose quality has a direct influence on various application scenarios, such as autonomous driving, robotic manipulation [25], surgery navigation [9], augmented reality and 3D reconstruction.

Thanks to advanced deep learning techniques, the performance of depth estimation methods has improved significantly over the last few years. Most previous work relied on ground-truth depth data and considered deep architectures for generating depth maps in a supervised manner [13, 24]. However, collecting vast and varied training datasets with accurate per-pixel ground truth depth data for supervised learning is a formidable challenge. To overcome this limitation, some recent works have shown that self-supervised methods are instead able to effectively tackle the depth estimation task [19, 27]. The approaches proposed in [7, 11] are particularly inspirational, where they took view synthesis as

a supervisory signal to train the network and exploited differences between the original input and synthesized view as penalties (*i.e.* a photometric image reconstruction cost and a disparity smoothness cost) to force the system to generate accurate disparity maps. However, although some works have tried to emphasize the complementary information in the stereo image pair and used shared weights when extracting features from input images [2, 19], the contextual information between the multiple views — especially some strong feature matches have not been effectively explored and exploited.

In this paper, we use an unsupervised learning setting and introduce the H-Net, an end-to-end trainable network for depth estimation given rectified stereo image pairs. The proposed H-Net effectively fuses information in the stereo pairs and combines epipolar geometry with learning-based depth estimation approaches. In summary, our main contributions in this paper are:

- A Siamese encoder-Siamese decoder network architecture in self-supervised learning schema was proposed, which fuses the complementary information in the stereo image pairs while enhancing the communication between them.
- A new mutual epipolar attention module was proposed to enforce the epipolar constraints in feature matching and emphasize the strong relationship between the features located along the same epipolar lines in rectified stereo image pairs.
- An optimal transport algorithm was explored and applied on the mutual epipolar attention module to make a further enhancement and incorporate semantic information in a novel fashion while filtering out outlier feature correspondences.

We demonstrate the effectiveness of our approach on the challenging KITTI [5] and Cityscapes datasets [3]. An ablation study was conducted by turning various components of the model off in turn, indicating the respective positive influence of each proposed module on the overall performance.

2. Related work

Estimating depth maps from stereo images has been explored for decades [1]. Accurate stereo depth estimation plays a critical role in perceiving the 3D geometric configuration of scenes and facilitating a variety of real world computer vision applications [12]. Recent work has shown that depth estimation from a stereo image pair can be effectively tackled by learning-based methods with convolutional neural networks (CNNs) [2].

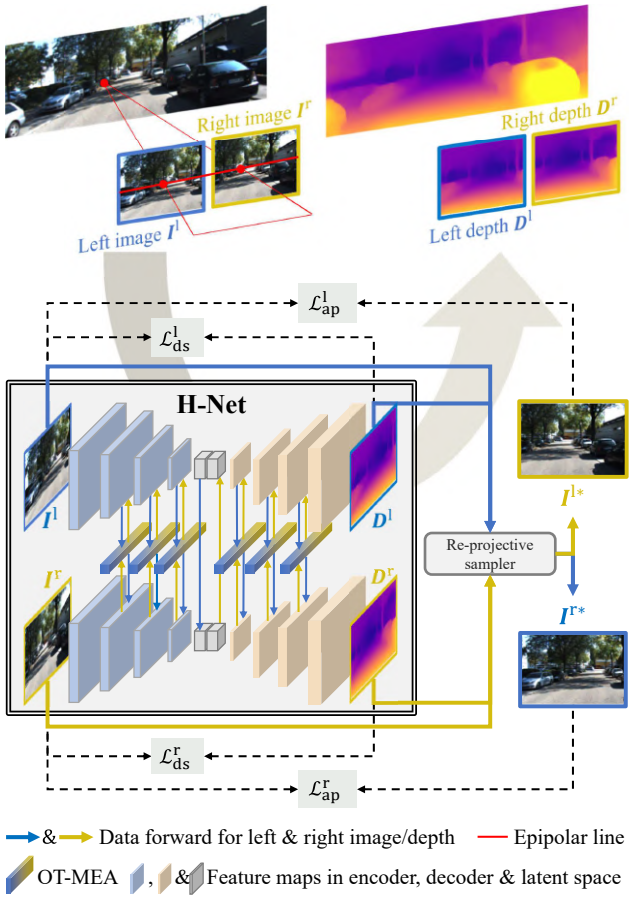

Figure 1. The network architecture of the proposed H-Net with new Optimal Transport based Mutual Epipolar Attention module (OT-MEA, as shown in Fig. 2) and self-supervised training details are shown in Sec. 3.4. Siamese encoder-decoder architecture with shared weights was used to extract features which are fed to the OT-MEA modules for exploring long-range dependencies of the epipolar geometry between stereo image pairs.

Due to the lack of per-pixel ground truth depth data, much work has investigated self-supervised depth estimation, where corresponding image reconstruction accuracy forms the supervisory signal during training [11]. It has been also shown that training with an added binocular color image helps single image depth estimation without requiring ground truth [7]. Andrea *et al.* [19] showed that the depth estimation results could be effectively improved within an adversarial learning framework, with a deep generative network that learned to predict the disparity map for a calibrated stereo camera using a wrapping operation. A pyramid stereo matching network was proposed in [2], where spatial pyramid pooling and dilated convolution were adopted to enlarge the receptive fields, while a stacked hourglass CNN was designed to further boost the utilization of

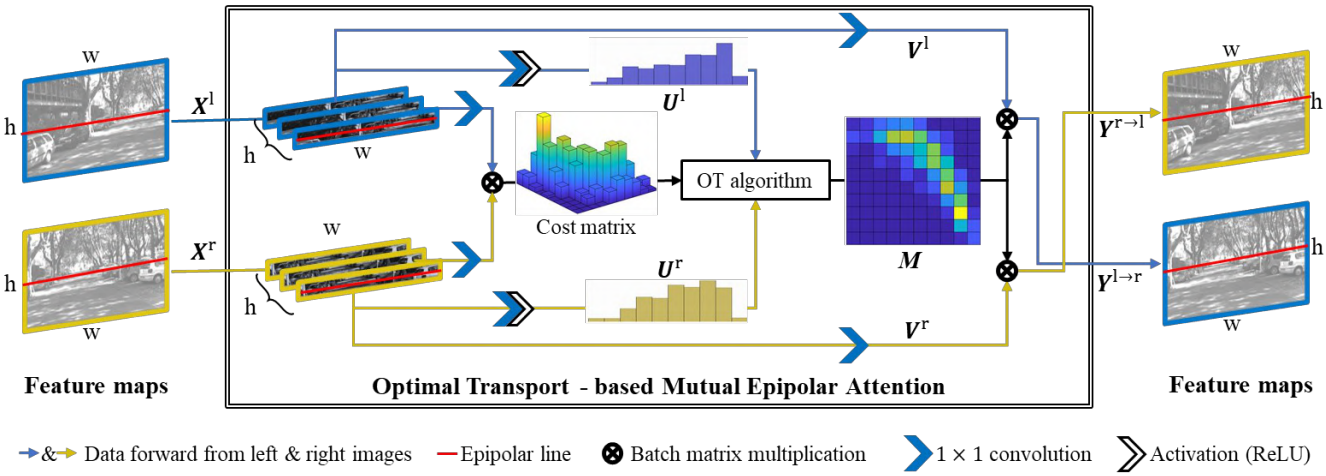

Figure 2. The Optimal Transport based Mutual Epipolar Attention block (OT-MEA) computed the correspondence between pixels from stereo image pairs (Eq. 1) while leveraging OT algorithm (Eq. 4) to assign varying weights for retrieval.

global context information. Duggal *et al*. [4] proposed a differentiable PatchMatch module to abandon most disparities without requiring full cost volume evaluation, and thus the specific range to prune for each pixel could be learned. Kusupati *et al*. [15] improved the depth quality by leveraging the predicted normal maps and a normal estimation model, and proposed a new consistency loss to refine the depths from depth/normal pairs.

In the multi-view (stereo) depth estimation task, it is natural to employ complementary features from different views to establish the geometric correspondences. Zhou *et al*. [28] presented a framework that learned stereo matching costs without human supervision by updating the network parameters in an iterative manner and guided by a left-right check. Joung *et al*. [12] proposed a framework to compute matching cost in an unsupervised setting, where the putative positive samples in every training iteration were selected by exploiting the correspondence consistency between two stereo images. Although these methods tried to explore the feature relationship between the stereo images, the concrete matching matrices were not effectively exploited or applied to the learning procedure, which leads to the loss of details and geometric information, especially the strong constraints on the epipolar line.

3. Method

3.1. H-Net architecture

In this paper, the encoder-decoder structure Monodepth2 [7] was adopted as the fundamental backbone, based on the U-Net [20]. As shown in Fig. 1, the proposed architecture consisted of a double-branch encoder and a double-branch decoder. To make the network compact and more suitable for self-supervised stereo training, inspired by [2] and [19], a Siamese Encoder - Siamese Decoder (SE-SD) structure was designed with shared weights between the two branches in both the encoder and the decoder which enabled the extraction of mutual information from the pair of input images. The Siamese Encoder (SE) of H-Net included two branches of Resnet18 [8] with shared trainable parameters. The left and right rectified images $\boldsymbol{I}^l, \boldsymbol{I}^r \in \mathbb{R}^{3\times h_0\times w_0}$ were fed into each branch of the SE to extract common features from the input images, where h_0, w_0 denote the image size. The outputs of the three deeper Residual-down-sampling (Res-down) blocks in the SE were interconnected with a novel mutual attention block proposed in this work — the so-called Optimal Transport-based Mutual Epipolar Attention (OT-MEA) block, shown in Fig. 2 and explained in detail below.

The abstract latent features from the encoder were fused in the middle part by concatenating the feature maps extracted from each SE block between the two branches. Each concatenated map was then convolved by two separate convolution layers with different trainable parameters.

The decoder took the fused latent features as inputs and generated sigmoid outputs for each input image similar to [6] and [7]. It was composed of the same number of Residual-up-sampling (Res-up) blocks as Res-down to recover the full resolution, as well as OT-MEA blocks inserted in the first three Res-up blocks. Each sigmoid output Ω of the decoder was transformed to scene depth as $D = 1/(a\Omega + b)$. The parameters a and b were selected to constrain depth D between 0.1 and 100 units.

3.2. Mutual Epipolar Attention

Here we introduce a mutual attention mechanism to give more emphasis to feature correspondences which lie on the same epipolar line.

Recently, Wang et al. [23] proposed the Non-Local (NL) block which allowed them to exploit global attention in an image sequence. This was then extended with the introduction of the Mutual NL (MNL) block [26] to explore the mutual relationships between different inputs in multi-view vision. However, global-range feature matching in the NL and MNL blocks suffers from the high number of parameters, memory requirement and training time. Furthermore, these blocks can be misled by repeated textures in the scenes.

To overcome the above limitations, we designed the Mutual Epipolar Attention (MEA) module to constrain feature correspondences to the same epipolar line between a pair of rectified stereo images. MEA was defined as:

$$\begin{cases} \boldsymbol{Y}^{\mathrm{l}\to\mathrm{r}} := \Psi(\boldsymbol{X}^{\mathrm{l}}) \otimes \Phi(\boldsymbol{X}^{\mathrm{l}}, \boldsymbol{X}^{\mathrm{r}}) \\ \boldsymbol{Y}^{\mathrm{r}\to\mathrm{l}} := \Psi(\boldsymbol{X}^{\mathrm{r}}) \otimes \Phi(\boldsymbol{X}^{\mathrm{r}}, \boldsymbol{X}^{\mathrm{l}}) \end{cases} \tag{1}$$

where $\otimes$ denotes the batch matrix multiplication, $\boldsymbol{X}^{\mathrm{l}}, \boldsymbol{X}^{\mathrm{r}} \in \mathbb{R}^{h\times c\times w}$ denote the transported and reshaped input signals from the two branches, and $\boldsymbol{Y}^{\mathrm{l}\to\mathrm{r}}, \boldsymbol{Y}^{\mathrm{r}\to\mathrm{l}} \in \mathbb{R}^{h\times c\times w}$ are the output signals from the MEA block. $\Phi : \mathbb{R}^{h\times c\times w} \times \mathbb{R}^{h\times c\times w} \to \mathbb{R}^{h\times w\times w}, (\boldsymbol{X}^1, \boldsymbol{X}^2) \mapsto \boldsymbol{M}^{1\to 2}$ is a pair-wise matching function — the so called retrieval function — which evaluates the compatibility between the two inputs. $\Psi : \mathbb{R}^{h\times c\times w} \to \mathbb{R}^{h\times c\times w}, \boldsymbol{X} \mapsto \boldsymbol{V}$ is a unary function which maps vectors from one feature space to another and is essential for fusion.

Following the settings in [23], the Embedded Gaussian (EG) similarity representation was used to define our matching function:

$$\Phi_{\mathrm{EG}}(\boldsymbol{X}^1, \boldsymbol{X}^2) := \mathrm{softmax}(\mathcal{C}_1(\boldsymbol{X}^1)^\top \otimes \mathcal{C}_2(\boldsymbol{X}^2)) \tag{2}$$

where $\mathcal{C}$ is the 1×1 convolution, and was also used in the unary function for vector mapping:

$$\Psi := \mathcal{C} \tag{3}$$

In the experimental work, the EG-based MEA and MNL modules were compared and denoted as EG-MEA and EG-MNL, respectively.

3.3. Optimal transport based mutual attention

In stereo vision, input images are captured from cameras at different positions and view angles resulting in slightly different fields of view. This can cause outliers in depth estimation due to incorrect feature correspondences in the areas which are not visible to both cameras. To eliminate outliers in these areas, the MEA module was enhanced to suppress the contribution of correspondences in these occluded areas during feature matching. The EG similarity representation defined in Eq.(2) cannot achieve this because all the areas of the input signals are equally considered.

For this purpose, we formulated the matching task in Eq.(1) as an optimal transport (OT) problem, as it has already been proven that OT improves semantic correspondence [16]. Thus, a new OT-based retrieval function is further proposed, tailored to our stereo depth estimation problem:

$$\begin{aligned} &\Phi_{\mathrm{OT}}(\boldsymbol{X}^1, \boldsymbol{X}^2) := \arg\min_{\boldsymbol{M}} \|\boldsymbol{M} \odot \mathrm{e}^{1-\mathcal{C}_1'(\boldsymbol{X}^1)^\top \otimes \mathcal{C}_2'(\boldsymbol{X}^2)}\|_1 \\ &\text{s.t.} \qquad \mathbf{u} \otimes \boldsymbol{M} = \Theta(\boldsymbol{X}^2),\ \mathbf{u} \otimes \boldsymbol{M}^\top = \Theta(\boldsymbol{X}^1) \end{aligned} \tag{4}$$

where $\odot$ denotes a Hadamard product, $\mathcal{C}'$ is a sequence operation of convolution and channel-wise Euclidean normalization, $\mathbf{u} \in \{1\}^{h\times 1\times w}$ is a matrix with all elements equal to 1. $\Theta : \mathbb{R}^{h\times c\times w} \to \mathbb{R}^{h\times 1\times w}, \boldsymbol{X} \mapsto \boldsymbol{U}$ is the sequence operation of convolution, ReLU activation and pixel-wise L1-normalization to generate the transported mass of pixels $\boldsymbol{U}$. The matrix $\boldsymbol{M}$ is the variable to be optimised and represents the optimal matching matrix $\boldsymbol{M}^{1\to 2}$.

Here, OT-based matching in Eq. (4) assigns to each pixel the sum of each column of the similarity weights in matching matrix $\boldsymbol{M}^{1\to 2}$, which is constrained by the mass:

$$\begin{cases} U^1_{ij} = \sum_k M^{1\to 2}_{ijk} \\ U^2_{ik} = \sum_j M^{1\to 2}_{ijk} \end{cases}, \forall i, j, k \in \mathbb{Z}, i \le h, j, k \le w \tag{5}$$

where U^1_{ij}, U^2_{ik} and $M^{1\to 2}_{ijk}$ are the elements of the $\boldsymbol{U}^1$, $\boldsymbol{U}^2$ and $\boldsymbol{M}^{1\to 2}$ respectively indexed by i, j, k. In contrast to the equal consideration by EG-based matching in Eq. (2), varying weights are assigned to different correspondences in Eq. (5), determined by the latent semantic messages forwarded from the input signals. This enables the OT module to suppress the outliers and focus on correspondences with more mass which lie on the semantic areas.

Since Eq. (4) is a convex optimization problem, the Sinkhorn algorithm was used to obtain the numerical solution of this OT problem [16]. OT matching based MEA is denoted as OT-MEA and Fig. 2 illustrates the implementation sketch of the OT-MEA used in H-net. Both MEA and OT modules can be used separately or in combination and we present their impact with an ablation study in Section 5.2. OT-MEA was also compared in our experimental work to the OT matching based MNL (OT-MNL).

3.4. Self-Supervised Training

For the left and right input images $\boldsymbol{I}^{\mathrm{l}}, \boldsymbol{I}^{\mathrm{r}} \in \mathbb{R}^{3\times h_0\times w_0}$, the sigmoid outputs of the H-Net were transformed to depth maps $\boldsymbol{D}^{\mathrm{l}}, \boldsymbol{D}^{\mathrm{r}} \in \mathbb{R}^{1\times h_0\times w_0}$ as explained in Section 3.1. By combining one of the depth maps (e.g $\boldsymbol{D}^{\mathrm{l}}$) and the counterpart input image ($\boldsymbol{I}^{\mathrm{r}}$), we were able to reconstruct the initial image ($\boldsymbol{I}^{\mathrm{l}*}$) using the re-projection sampler [10]. Here we used the left image $\boldsymbol{I}^{\mathrm{l}}$ as an example to present the supervisory signal and loss components. The final loss function included the loss terms for both left and right images. The similarity between the input image $\boldsymbol{I}^{\mathrm{l}}$ and the reconstructed

image $\boldsymbol{I}^{\mathrm{l}*}$ provides our supervisory signal. A photometric error function $\mathcal{L}_{ap}^{l}$ was defined as the combination of L_1-norm and structural similarity index (SSIM) [7]:

$$\mathcal{L}_{\mathrm{ap}}^{\mathrm{l}} = \frac{1}{N}\sum_{i,j}\frac{\gamma}{2}(1-\mathrm{SSIM}(I_{ij}^{\mathrm{l}}, I_{ij}^{l*}))+(1-\gamma)\|I_{ij}^{\mathrm{l}} - I_{ij}^{l*}\|_1 \quad (6)$$

where, N denotes the number of pixels and γ is the weighting for L_1-norm loss term. To improve the predictions around object boundaries, an edge-aware smoothness term $\mathcal{L}_{ds}$ was applied [7, 11]:

$$\mathcal{L}_{\mathrm{ds}}^{\mathrm{l}} = \frac{1}{N}\sum_{i,j}|\partial_x(d_{ij}^{\mathrm{l}*})|e^{-|\partial_x I_{ij}^{\mathrm{l}}|} + |\partial_y(d_{ij}^{\mathrm{l}*})|e^{-|\partial_y I_{ij}^{\mathrm{l}}|} \quad (7)$$

where $d^{l*} = d^{l}\sqrt{d^{l}}$ represents the mean-normalized inverse of depth ($1/\boldsymbol{D}$) which aims at preventing shrinking of the depth prediction [22].

To overcome the gradient locality of the re-projection sampler, we adopted the multi-scale estimation method presented in [7], which first upsamples the low resolution depth maps (from the intermediate layers) to the input image resolution and then reprojects and resamples them. The errors were computed at the higher input resolution. Finally, the photometric loss and per-pixel smoothness loss were balanced by the smoothness term λ and the total loss was averaged over each scale (s), branch (left and right) and batch:

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} &= \frac{1}{2m}\sum_{s=1}^{m}(\mathcal{L}_s^{\mathrm{l}} + \mathcal{L}_s^{\mathrm{r}}) \\ &= \frac{1}{2m}\sum_{s=1}^{m}\left((\mathcal{L}_{\mathrm{ap}}^{\mathrm{l}} + \lambda\mathcal{L}_{\mathrm{ds}}^{\mathrm{l}}) + (\mathcal{L}_{\mathrm{ap}}^{\mathrm{r}} + \lambda\mathcal{L}_{\mathrm{ds}}^{\mathrm{r}})\right)\end{aligned} \quad (8)$$

4. Experiments

We trained and evaluated H-Net on KITTI2015 [5], where there were 22600 pairs for training and 888 for validation. The same intrinsics were used for all images. The principal point of the camera was set to the image center and the focal length was the average of all the focal lengths in KITTI. All of the images were rectified and the transformation between the two stereo images was set to be a pure horizontal translation of fixed length. During the evaluation, only depths up to a fixed range of 80m were evaluated per standard practice [7]. As our backbone model, we used Monodepth2 [7] and kept the original ResNet18 [8] as the encoder. Furthermore, we also trained and tested H-Net on the Cityscapes dataset [3] to verify its generalisability.

We compared our results with other supervised and self-supervised approaches and both qualitative and quantitative results were generated for comparison. The aim was to prove that the proposed modules were able to benefit the self-supervised depth estimation performance. Hence, to better understand how each component influenced the overall performance, an ablation study turned various components of the model off in turn.

4.1. Implementation Details

The H-Net was trained using the PyTorch library [17], with an input/output resolution of 640×192 and a batch size of 8. The L_1-norm loss term γ was set to 0.85 and the smoothness term λ was 0.001, which were determined by experiments and were consistent with related methods [7]. The number of scales m was set to 4, which meant that there were 4 output scales in total with resolutions $\frac{1}{2^0}$, $\frac{1}{2^1}$, $\frac{1}{2^2}$ and $\frac{1}{2^3}$ of the input resolution. The model was trained for 20 epochs using the Adam optimizer [14] requiring approximately 14 hours on a single NVIDIA 2080Ti GPU. The learning rate was set to 10^{-4} for the first 15 epochs and dropped to 10^{-5} for the remainder. As with previous papers [7], a Resnet encoder with pre-trained weights on ImageNet [21] proved able to improve the overall accuracy of the depth estimation and to reduce the training time [7, 11].

5. Results and Discussion

5.1. KITTI Results

The qualitative and quantitative results on the KITTI 2015 [5] are shown in Table 1 and Figure 3. In Table 1, it can be seen that the proposed H-Net provides a superior overall performance, which indicates that our model can learn from the geometric constraints and benefits from the optimal transport solution. To prove that the improvements were not just from the stereo input, we modified the input of the Monodepth2 [7] to concatenate stereo images with the rest settings unchanged. From the fourth row it is clear that although the stereo input benefited Monodepth2, the performance was still not as good as H-Net. For the quantitative results the depth maps generated by our model contained more details, *i.e.* the structural characteristics of buildings, bushes, and trees.

5.2. KITTI Ablation Study Results

The results of ablation study on the KITTI dataset are shown in Table 2. The impact of the Siamese encoder- Siamese decoder (SE-SD), mutual epipolar attention (MEA) and optimal transport (OT) were evaluated. The backbone Monodepth2 model [7] performed the worst without these contributions, but by changing the architecture to a Siamese encoder - Siamese decoder, the evaluation measures steadily improved. The reason might be that fusing the complementary information between the stereo image pair gave the framework a higher chance to generate accurate

Table 1. Quantitative results. Comparison of our proposed H-Net to existing methods on KITTI2015 [5] using the Eigen split unless marked with 'Full Eigen', which indicates the full Eigen dataset. The best result in each category are presented in bold while the second best results are underlined. Metrics labeled by red mean *lower is better* while labeled by blue mean *higher is better*. S:Stereo; M:Mono.

Method	Train	Infer	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Pilzer [19]	S	S	0.152	1.388	6.016	0.247	0.789	0.918	0.965
PFN [18]	S	S	0.102	0.802	4.657	0.196	0.882	0.953	0.977
Monodepth2 [7] (backbone)	S	M	0.109	0.873	4.960	0.209	0.864	0.948	0.975
Monodepth2 [7] (Concat)	S	S	0.082	0.752	4.407	0.183	0.914	0.960	0.978
H-Net (Ours)	S	S	**0.076**	**0.607**	**4.025**	**0.166**	**0.918**	**0.966**	**0.982**

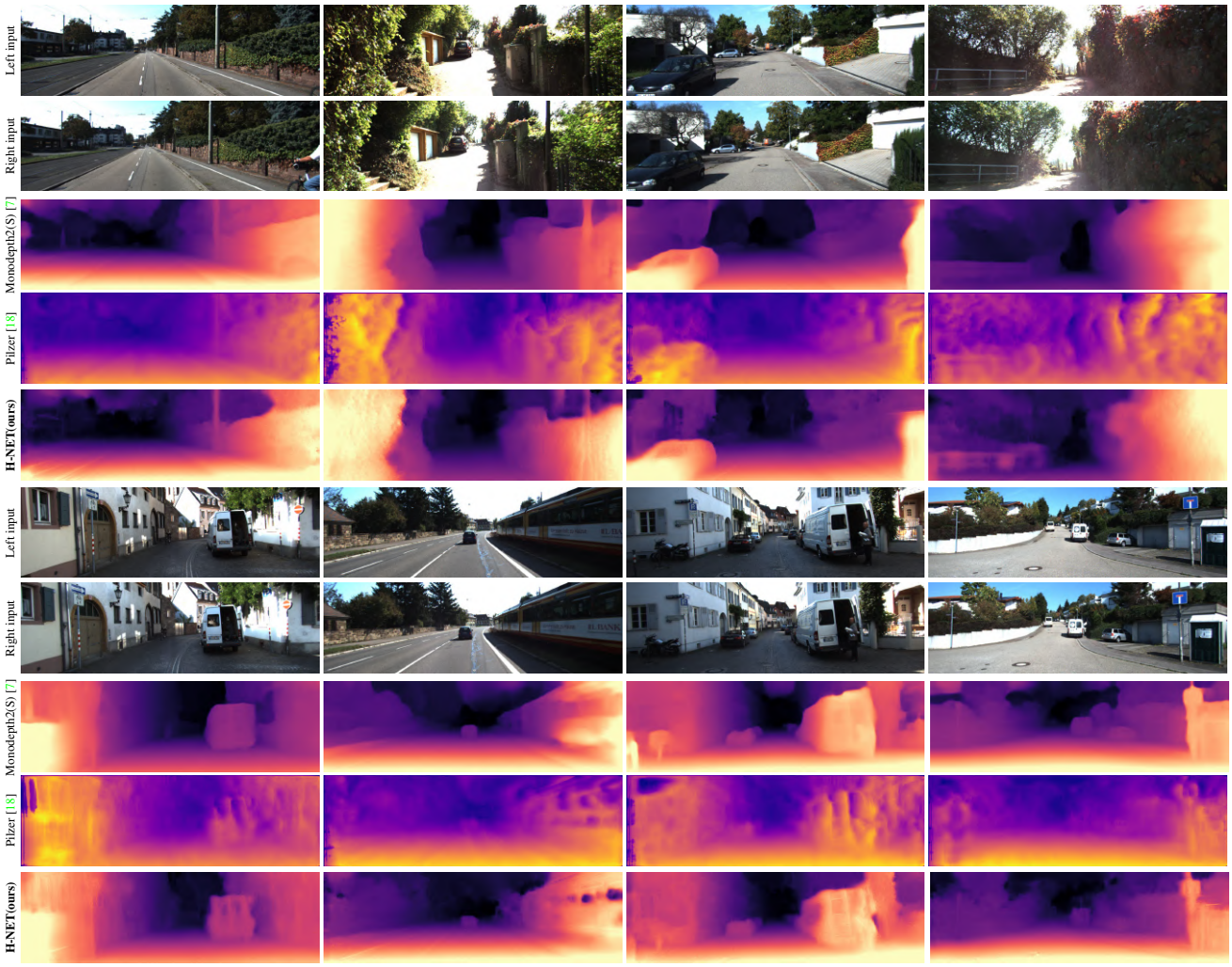

Figure 3. Qualitative results on the KITTI Eigen split. The depth prediction are all for the left input image. Our H-Net in the last row generates the depth maps with more details and performs better on distinguishing different parts in one object, *i.e.* buildings, kerbs bushes and trees, which reflects the superior quantitative results in Table 1.

predicted depth maps, especially in self-supervised training setting.

The MEA and OT modules were all incorporated in the SE-SD architecture. The comparison between Row 2 and Row 4 shows that the MEA module benefits depth estimation performance in all the evaluation measures, especially on metrics that are sensitive to large depth errors *e.g.* RMSE. The significantly large improvement of the SE-SD architecture with MEA is likely due to the epipolar constraint, which allowed the network to learn strong corre-

Table 2. Ablation Study for different variants of H-Net on KITTI2015 [5] using full Eigen dataset with comparison to our backbone Monodepth2 [7]. We evaluate the impact of the Siamese encoder- Siamese decoder (SE-SD), mutual epipolar attention (MEA) and optimal transport (OT). Metrics labeled by red mean *lower is better* while labeled by blue mean *higher is better*

Setting	SE-SD	MEA	OT	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [7]	✗	✗	✗	0.109	0.873	4.960	0.209	0.864	0.948	0.975
SE-SD	✓	✗	✗	0.096	0.700	4.403	0.189	0.894	0.960	0.979
SE-SD w/ EG-MNL	✓	✗	✗	0.086	0.701	4.289	0.178	0.912	0.964	0.980
SE-SD w/ EG-MEA	✓	✓	✗	0.080	0.665	4.086	0.173	0.917	0.964	0.981
SE-SD w/ OT-MNL	✓	✗	✓	0.082	0.725	4.279	0.180	0.917	0.962	0.979
H-Net (Ours)	✓	✓	✓	**0.076**	**0.607**	**4.025**	**0.166**	**0.918**	**0.966**	**0.982**

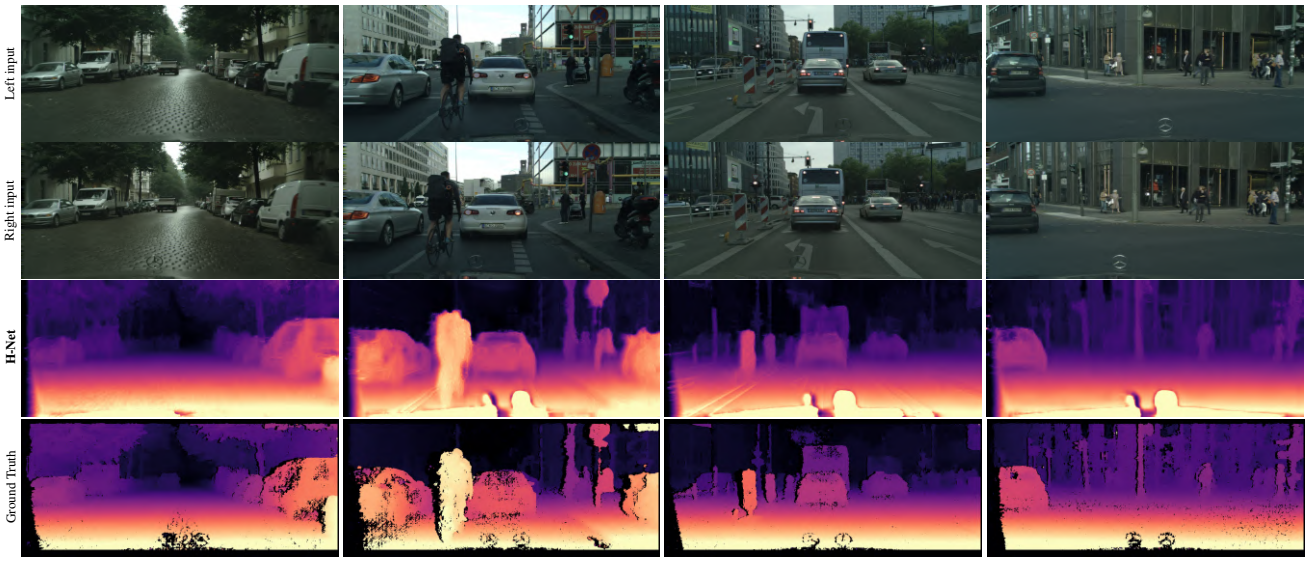

Figure 4. Qualitative results on the Cityscapes dataset. Our H-Net generates very close predictions compared with the ground truth.

spondences limited to the same epipolar lines in the rectified stereo images. The impact of the OT-MNL is presented in Row 5, where there is a dramatic increase in most evaluation metrics compared to the SE-SD in Row 2. The reason might be that the optimal transport algorithm further improved the MEA by increasing the correct correspondence weights, merging the semantic features while suppressing outliers. In the last row, by combining the backbone with all of our components, the effectiveness of the final framework was significantly improved, as expected, and state-of-the-art results were observed. Besides, although our OT-MEA module was inspired by the MNL, our results outperformed the same SE-SD architecture with MNL.

The number of parameters for each of the examined settings was also estimated. While all of our proposed components contributed to the overall performance in the self-supervised depth estimation task, the number of parameters was barely increased. Table 3 shows that our OT-MEA module costs 0.6 million (2.0%) additional parameters compared with the pure SE-SD architecture.

Table 3. Number of Parameters (M:million) for our models with different settings of the mutual attention module.

Setting	Num of Parameters
SE-SD (baseline)	30.7M
EG-MNL	+0.3M(1%)
EG-MEA	+0.6M(2%)
OT-MNL	+0.3M(1%)
OT-MEA (Ours)	+0.6M(2%)

5.3. Cityscapes results

The performance of H-Net was further evaluated on the Cityscape dataset. The results in Figure 4 show the accuracy of the depth estimated by H-Net compared to the ground truth, with detailed reconstructions of objects such as cars, humans, and trees.

6. Conclusion

In this paper we presented a novel network, the H-Net, for self-supervised depth estimation. By designing the Siamese encoder-Siamese decoder architecture, exploiting the mutual epipolar attention, and formulating the optimal transport problem, the global-range correspondence between stereo image pairs and the strongly related feature correspondences satisfying an epipolar constraint were explored and fused. This was shown to benefit the overall performance on public datasets and gave a large improvement in evaluation measures, indicating that the model effectively overcame the limits of other self-supervised depth estimation methods.

Acknowledgment

This work was supported by the UK National Institute for Health Research (NIHR) Invention for Innovation Award NIHR200035, the Cancer Research UK Imperial Centre, the Royal Society (UF140290) and the NIHR Imperial Biomedical Research Centre.

References

[1] Stephen T Barnard and Martin A Fischler. Computational stereo. *CSUR*, 1982. 2

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018. 2, 3

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 5

[4] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *ICCV*, 2019. 3

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 5, 6, 7

[6] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, July 2017. 3

[7] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 1, 2, 3, 5, 6, 7

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[9] Baoru Huang, Jian-Qing Zheng, Anh Nguyen, David Tuch, Kunal Vyas, Stamatia Giannarou, and Daniel S Elson. Self-supervised generative adversarial network for depth estimation in laparoscopic images. In *MICCAI*, pages 227–237. Springer, 2021. 1

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NeurIPS*, 28, 2015. 4

[11] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *CVPR*, 2020. 1, 2, 5

[12] Sunghun Joung, Seungryong Kim, Kihong Park, and Kwanghoon Sohn. Unsupervised stereo matching using confidential correspondence consistency. *TITS*, 2019. 2, 3

[13] 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 *ICCV*, 2017. 1

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[15] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *CVPR*, 2020. 3

[16] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *CVPR*, 2020. 4

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[18] Andrea Pilzer, Stéphane Lathuilière, Dan Xu, Mihai Marian Puscas, Elisa Ricci, and Nicu Sebe. Progressive fusion for unsupervised binocular depth estimation using cycled networks. *PAMI*, 2019. 6

[19] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *3DV*, 2018. 1, 2, 3, 6

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[22] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018. 5

[23] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 4

[24] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, 2020. 1

[25] Dandan Zhang, Frank P-W Lo, Jian-Qing Zheng, Wenjia Bai, Guang-Zhong Yang, and Benny Lo. Data-driven microscopic pose and depth estimation for optical microrobot manipulation. *ACS Photonics*, 7(11):3003–3014, 2020. 1

[26] Jian-Qing Zheng, Ngee Han Lim, and Bartłomiej W Papież. D-net: Siamese based network for arbitrarily oriented volume alignment. In *International Workshop on Shape in Medical Imaging*, pages 73–84. Springer, 2020. 4

[27] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017. 1

[28] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *ICCV*, 2017. 3