Fall 11-24-2021

ChiTransformer: Towards Reliable Stereo from Cues

Qing Su

ChiTransformer: Towards Reliable Stereo from Cues

by

Qing Su

Under the Direction of Jonathan Shihao Ji, PhD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2021

ABSTRACT

Current stereo matching techniques are challenged by restricted searching space, occluded regions and sheer size. While monocular depth estimation is spared from these challenges and can achieve satisfactory results with monocular cues, the lack of stereoscopic relationship renders the monocular prediction less reliable on its own especially in highly dynamic or cluttered environments. To address these issues in both scenarios, an optic-chiasm-inspired self-supervised binocular depth estimation method is proposed in thesis, wherein vision transformer with gated positional cross-attention layer is designed to enable feature-sensitive pattern retrieval between views, while retaining the extensive context information aggregated through self-attentions. This crossover design is biologically analogous to the optic-chasma structure in human visual system and hence the name, ChiTransformer. It leverages strengths of both monocular and binocular approaches. Our experiments show this architecture yields substantial improvements on self-supervised stereo approaches by 15% and can be used on both rectilinear images and fisheye images.

INDEX WORDS: ChiTransformer, Spectral polarization, Attention, Large disparity, Occluded estimation, Learned epipolar geometry, Self-supervised training

ChiTransformer: Towards Reliable Stereo from Cues

by

Qing Su

Committee Chair: Jonathan Shihao Ji

Committee: Rolando Estrada

Saeid Belkasim

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

December 2021

DEDICATION

This thesis is genuinely and wholeheartedly dedicated to my beloved parents, who have always been my inspiration and given me strength throughout the whole research journey. Their constant support in emotion, spirit and finance is truly appreciated

To my precious friend, Eric Feng, who always stands by me, supports me and pulled me through the hardships with patience, kindness, and fortitude.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Jonathan Shihao Ji for his support of my MS study and research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CA	cross-attention
ChiT	ChiTransformer
CNN	convolutional neural network
DCR	depth cue rectification
GPCA	gated positional cross attention
MDE	monocular depth estimation
MHA	multi-headed attention
NLP	natural language processing
RSB	reassemble block
SA	self-attention
ViT	vision transformer

1 INTRODUCTION

Transformer has been proven to be very effective in various vision tasks. Recently, vision transformer is exploited to depth estimation and achieves superior results. In this paper, we push its capability further to stereo vision tasks and non-rectilinear settings.

In the context of computer vision, nowadays almost all mainstream depth estimation methods are deep learning based and can be roughly categorized into two prevalent methodologies, namely, stereo matching and monocular depth estimation. Stereo matching has traditionally been the most investigated area due to its strong connection to the human visual system. The task is to find or estimate the correspondences of all the pixels of two rectified images [3, 4, 51]. Virtually, all the current works resort to convolutional neural network (CNN) based methods to calculate the matching cost since its first introduction to the task by [14,68] in 2015. Following the work of FlowNet [14], more than 150 papers have been published using CNN-related methods [36], pushing the performance forward by more than 50%. Some deep-seated issues such as thin structures, large texture-less areas, and occlusions have been mitigated or addressed [29, 70] over time. So far, stereo matching is the most adopted technique in majority of passive stereo applications.

However, the applications that entail depth estimation grow increasingly demanding as visual systems are greatly downsized and installed on platforms with higher mobility (e.g., UAV, commercial robots). This indicates a more congested, cluttered and dynamic operating environment where the once side issues become major ones, i.e., large disparity, large occlusion and non-rectilinear images might be involved. Most existing stereo matching methods are not set up for this new trend and hence fail to address these issues properly.

On the other hand, monocular depth estimation (MDE) is spared from these issues as depth is estimated from a single view. Current works, following [17], leverage deep learning models to derive more descriptive cues to achieve superior predictions. More recent works focus on fusing multi-scale information to further improve the pixel level depths [37, 41]. Lately, vision transformer is exploited in the task and yields globally organized and coherent predictions with finer granularity [5,48].

State-of-the-art MDE methods can achieve impressive results with relative accuracy $\delta 3$ surpasses 0.99 by supervised training [19, 24, 40, 66]. However, the reliability of MDE estimation is essentially based on the assumption that the scenes in the real world are mostly regular. Therefore, due to the lack of stereo relation, the MDE is more delimited to its dataset and susceptible to “unfamiliar” scenes. This renders the MDE alone not reliable in safety-critical applications such as autonomous driving and visual-aided UAV.

Up to this point, we can see that the limitations and advantages of stereo matching and MDE are complementary. Following this observation, we propose a novel method that jointly addresses their limitations by crossing over the stereo and depth estimation approaches such that stereo in- formation can be injected into the MDE process to rectify and improve the estimation made from depth cues.

We introduce ChiTransformer, an optic-chiasm-inspired dense prediction transformer. ChiTransformer adopts the recently published vision transformer (ViT) [13] as backbone and extends the encoder-only transformer to an encoder-decoder structure similar to the transformer models [12, 56] for natural language processing (NLP). Unlike the end-to-all connection in NLP transformer, ChiTransformer adopts an interleaved connection for cross-attention to progressively instill the encoded depth cues and contextual information from the nearby view to the master view in a self-regressive process. Our main contribution is the design of a retrieval cross-attention layer. Instead of attending and curing over multi-level contextual relations from the encodings like regular multi-head attention (MHA), the cross-attention mechanism of ChiTransformer aims to retrieve depth cues with strong contextual and feature coincidence from the other view. To achieve this, we condition the initial state (query) with a self-adjoint operator G without breaking the convergence rule of modern Hopfield network [47]. The positive-definite G is spectrally decomposed to enable polarized attention within the encoded feature space to emphasize on certain cues while preserving as much of the original information as possible. We show that this design facilitates reliable retrieval and leads to finer feature-consistent details on top of the globally coherent estimations. Moreover, the model can be further extended to non-rectilinear images such as fisheye by using gated positional embedding [10]. We model the epipolar geometry with learnable quadratic polynomials of relative positions. Considering the per-pixel labeled data is challenging to acquire at scale and let alone for the non-rectilinear images, we choose to train the model with self-supervised learning strategy tailored from the work in [23].

Experiments are conducted on depth estimation tasks that provide stereo pairs. Our result shows that ChiTransformer delivers significant improvement by more than 15% compared to the top-performing self-supervised stereo methods. The architecture is also tested on stereo tasks to evaluate the gain brought by stereo cues and the underlying reliability yielded by the instilled stereo information. To show the potential of ChiTransformer in non-rectilinear images, we train our model to predict the distances on the translated synthetic fisheye sequences from [16] and achieve visually satisfactory results.

In contrast to traditional stereo methods, our approach foregoes pixel-level matching optimization but leverages the context-infused depth cues of both images to improve the overall depth prediction accordingly. With a global receptive field, ChiTransformer is not restricted to certain epipolar geometry such as the horizontal collinear epipolar lines of rectified regular stereo pairs. It is also able to treat large disparity. Furthermore, with the inherent capability of depth estimation within a single image, estimation at large occluded area can be properly handled rather than being masked out, interpolated, or left untreated. Enhanced from current MDE methods, our approach provides reliable prediction with guided cues in stereo pair which makes ChiTransformer more suitable for complex and dynamic environments.

2 RELATED WORKS

Since the publication of [17] (2014) and [18] (2015), end-to-end trainable CNN-based models have been the prototypical architecture for dense depth [22, 23, 49] or dis- parity estimation [26, 28, 53, 64]. The principal idea is to leverage learned representation to improve matching cost [33,51] or depth cues [6] with appropriately large local contextual information. The prevalent encoder-decoder structure enables progressive down-sampling and up-sampling of representation at different scales [9, 15, 41, 63, 71]. Intermediate results from previous layers are often reused to recover fine-grained estimations while ensuring sufficiently large context.

After showing exemplary performance on a broad range of NLP tasks, attention and, in particular, transformer has demonstrated competitive or superior capability in vision tasks such as image recognition [13,54], object detection [8, 73], semantic segmentation [65], super-resolution [62], image impainting [69], image generation [50], text-image synthesis [1], etc. The successes also sparked interest in the community of stereo and depth estimation. [60] leveraged cascaded attentions to calculate the matching cost along the epipolar lines and achieved competitive results among self-supervised stereo matching methods [2, 31, 39, 72]. More recently, vision transformer was leveraged in place of convolution network as backbone for dense depth prediction in [48] and achieved significant improvement by 28% compared to the state-of-the-art convolutional counterparts. In [5], a mini-ViT block is employed in the refinement stage to facilitate the adaptive depth bin calculation. The work tops KITTI [21] and NYUv2 [52] leaderboards. Inspired by [48], our method leverages the capability of ViT in learning long range complex context information to help rectify depths cue instead of performing stereo matching.

Most works mentioned above are fully supervised, which necessitate pixel-wise labeled ground truth for training. However, it is challenging to acquire varied real-world setting at a large scale. One of the workarounds is to adopt self-supervised learning. For stereo training, usually pixel disparities of synchronized stereo pairs are predicted [2, 31, 39, 63, 72], while for monocular training, not only depth but also camera pose has to be estimated to help reconstruct the image and constrain the estimation network [7, 22, 57, 67, 72]. Considering the versatility and the potential application environment of our method, we choose self-supervised training for ChiTransformer. Details are discussed in section 3.

3 METHOD

This section introduces the overall architecture of the ChiTransformer with elaboration on the key building blocks. We follow the successful configuration of the vision transformer [13] as the backbone and maintain the prevalent overall encoder-decoder structure for their repeatedly verified success in various dense prediction tasks. We show the interplay of the encoded representations or cues between a stereo pair in ChiTransformer and how they can be effectively converted into dense depth predictions. The intuition for the elicitation and success of this method is discussed.

3.1 Architecture

3.1.1 Overview

The complete architecture is shown in Figure 1. ChiTransformer employs a pair of hybrid vision transformers as backbone with ResNet-50 [25] for stereo pair patch embedding. The parameters of the two ResNet tower are shared to ensure consistency in representation. The 2D-arranged patch embeddings are first projected to 768 dimensions, then flattened and summed with positional embeddings before fed into attention blocks. For an image of size $H \times W$, if the embedding size is (P, P), the result would be a set $T = \{t_0, ..., t_{N_p}\}$, where $N_p = H{\cdot}W/ P^2$ and t_0 is the class token. Here, patches are in the role of ‘words’ for transformer. We will address patches as “tokens” or use them interchangeably here- after. The attention block for the reference view closely follows the design in [13] with “classification token” included. Whereas master tokens are self-attended in the first multiple layers followed by cross-attention (CA) and self-attention (SA) layers in an interleaved fashion with CA layer being inserted after every other SA layer. The output tokens of the master ViT (and reference ViT in training) are then reassembled into an image-like arrangement. Feature representations Is at different scales s ∈ S are progressively aggregated and fused into the final depth estimation in the fusion block, which is modified by RefineNet [42]. Fusion block is shared for both views in training phase but dedicated to the master view in inference.

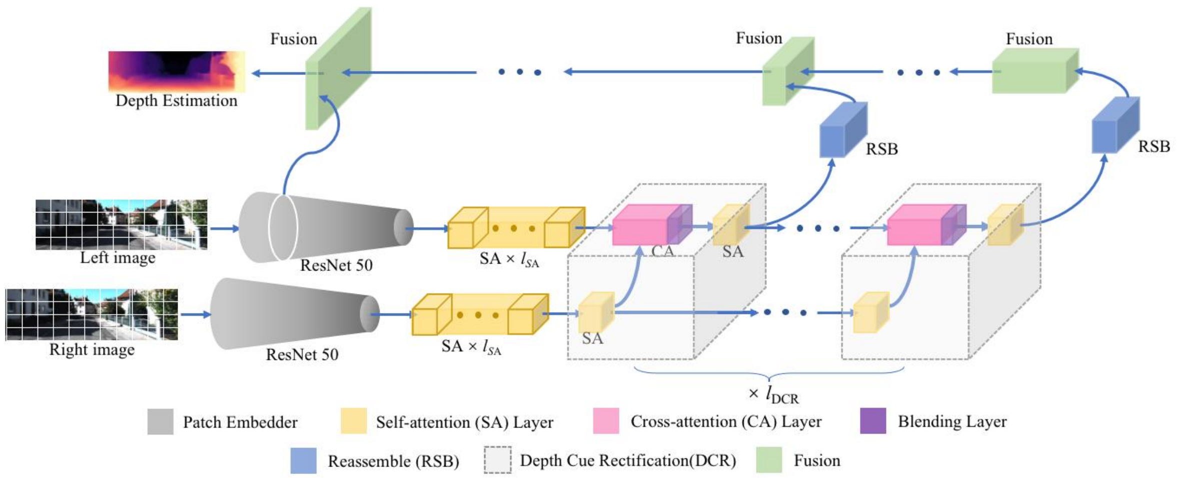


Figure 1. Architecture of ChiTransformer

A stereo pair (left:master, right:reference) is initially embedded into tokens through a Siamese ResNet-50 tower. The 2D-organized tokens from the two images are flattened and then augmented with learnable positional embeddings and an extra class token, respectively. Then tokens are fed into two self-attention (SA) stack of size l_{SA} *in parallel. After that, tokens are fed into a series* l_{DCR} *of depth rectification blocks (DCR) in each of which tokens of reference image go through an SA layer while tokens of master go through a polarized cross-attention (CA) layer, followed by an SA layer. In the polarized CA layer, relevant tokens from the output of reference SA are fetched to rectify the master's depth cues. Tokens from different stages are afterwards reassembled into an image-like arrangement at multiple resolution (blue) and progressively fused and up-sampled through fusion block to generate a fine-grained estimation.*

3.1.2 Attention layers

Self-attention layer is the crucial part for transformers and other attention-based methods to achieve superior performance over their non-attention competitors. The key advantage is that complex context information can be gathered in a global scope. With multilayer of SA, encodings get progressively tempered with the context information as it goes deeper into the attention layers. It is this mechanism that beget globally coherent predictions. There- fore, instead of putting immediate connection to the CA layer, we place multiple (l_{SA} = 4) SA layers at the output of the embedder. Cues with appropriate amount of context information result in more reliable pattern retrieval in the subsequent CA layers. This design improves both training convergence and prediction performance.

The *cross-attention* layer is our key contribution in Chi- Transformer. It is the enabler of the stereopsis through the fusion of high-level depth cue expressions from two views. We argue that the effectiveness of the traditional 4-step strategy would be largely weakened as the sources of ill-posedness such as occlusion, wider and closer range of depth, depth discontinuity and nonlinearity become increasingly frequent or prominent. Current deep learning- based methods rely on the learned rich representation to construct cost volume which is then regularized to make estimation. The output quality, in this case, largely de- pends on both the quality of the representation and the conformity of the scene to the matching regularizing assumptions [51]. While good representations can be learned with many approaches, there are few ways to fix up an impaired cost volume when scenes are far away from being appropriate for stereo matching. Therefore, instead of clinging to the matching strategy, we propose a novel pattern retrieval mechanism inspired by associative memory to retrieve the correspondent pattern from the other view. We assume that a set of patterns can be learned to separate well such that each pattern can be retrieved in at least meta-stable state, i.e., fixed average of similar patterns [47]. Modeled by modern Hopfield network [11, 35], the retrieval rule of the associative memory elegantly coincides with the attention mechanism of the transformer. Naturally, we lever- age cross-attention layer to retrieve patterns (tokens) from reference view to master view. To facilitate reliable effective retrieval, we devise a new attention mechanism – polarized attention, which enables feature sensitive retrieval while preserving the context information contained in the pattern without breaching the convergence rule. From [60], we observe that direct attention over representations at the output of CNN reduces to cosine similarity-based matching. Without position-dependent context information over extensive scope, patterns are liable to ill-posedness and low separability.

Given a retrieved token pair (${}^{m}t_i$, ${}^{m}t_i'$), $\forall i = 1, \cdots, N_p$ from the preceding CA layer and the class token pair (${}^{m}t_0$, ${}^{m}t'_0$), where m indicates master view, depth cues are then rectified through the following blending process:

$$f_{proj}({}^{m}t_i, {}^{m}t_i') = \mathrm{MLP}([{}^{m}t_i, {}^{m}t_i']), \tag{1}$$

$${}^{m}t_i = {}^{m}t_i + Heat(p\_a_i) \cdot LN(f_{proj}({}^{m}t_i, {}^{m}t_i')), \tag{2}$$

where GELU is used for MLP nonlinearity, LN is layer normalization, P_a_i is the vector of attention scores of ${}^{m}t_i$, and $Heat(\cdot)$ is the confidence score calculated with stabilized attention entropy,

$$H(p\_a_i) = -\sum_{k=1}^{N_p} p\_a_{i,k} \log(p\_a_{i,k} + \epsilon) \tag{3}$$

and

$$Heat(p\_a_i) = 1 - g(H(p\_a_i), \tau, c) \tag{4}$$

Heat is set to 1 for class token. Clamping function $g(\cdot)$ can be sigmoid or smooth-step function with temperature τ and offset c. By doing so, tokens retrieved back in fixed state, i.e., with a very low entropy, would be securely rectified whereas those with high entropy are inhibited from being updated as they are very likely to reside in occluded area. Thus, the depth in occluded or "uncertain" areas are left to the power of SA layers to speculate its value with context information and rectified cues from non-occluded areas.

3.1.3 Fusion block

Our convolutional decoder follows the refinement block in [42, 48]. The output of attention layers $\boldsymbol{t} \in \mathbb{R}^{(Np+1)\times D}$ is reassembled into an image-like arrangement $\mathbb{R}^{H'\times W'\times D'}$ through a four-step operation:

$$\mathrm{RSB} = (\mathrm{rescale} \circ \mathrm{reshape} \circ \mathrm{MLP} \circ \mathrm{cat}) \tag{5}$$

The class token is concatenated with all other tokens before being projected to dimension D′ to get $\boldsymbol{t}_{\backslash 0}$. Then it is reshaped into 2D shape as per the original arrangement of the image embedding. Finally, $\mathbf{t}_{\backslash 0}$ is re-sampled to size ${}^{H}/_{P} \cdot S_l \times {}^{W}/_{P} \cdot S_l \times D_l$ for different scales at level l. Re-sampling method is 2D transposed convolution (deconv) for $S_l > 1$, and strided 2D convolution for $S_l < 1$. For our model, features from level $l_{attn} = \{12, 8, 4\}$ in attention block and level $l_{res} = \{1, 0\}$ in ResNet50 are reassembled. The re- assembled feature maps from consecutive levels are finally fused through customized feature fusion block from RefineNet [42]. At each level, feature map is up-sampled by a factor of 2 and finally reaches half the resolution of the input images.

The architecture of ChiTransformer is structurally similar and biologically analogous to the *optic-chiasma* structure in our visual system (figure 2), where visual field covered by both eyes is fused to enable the processing of binocular depth perception by stereopsis [51], hence the name of our model.

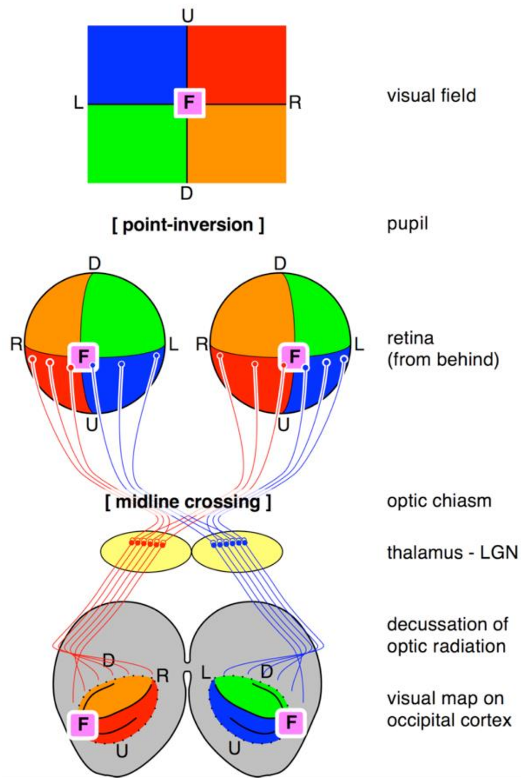


Figure 2. Optic-chiasma. Transformations of the visual field toward the visual map on the primary visual cortex in vertebrates. U=up; D=down; L=left; R=right; F=fovea

3.2 Polarized Attention

We propose a new attention mechanism to highlight or suppress features, which is very much like signal polarization but in channel domain. Ideally, for a set of tokens represented in tensor $\boldsymbol{t} = (t_1, \cdots, t_N)$ that is well separated, highlighting or suppressing can be potentially achieved in token-wise granularity. However, ideal separability is hard to achieve in practice and the attention tensor A for regular attention mechanism is calculated as,

$$\boldsymbol{A} = softmax(\beta t^{\top} W_t^{\top} W_{\xi} \xi), \tag{6}$$

which is prone to be noisy with joint activation over all channels and hard to learn directly for *W*'s. While the prevalent MHA seeks for multi-level context instead of retrieval as tokens are mapped to different (sub-)spaces for each head that generates its own attention weights and output with the projected tokens. However, to enforce the retrieval behavior with MHA by aggregating attention weights of all heads before output calculation simply reduces MHA to a regular attention. Therefore, without loss of generality, we stick to the Hopfield network update rule to ensure retrieval behavior and initialize the query pattern with a self-adjoint operator, $\mathbf{G} \in \mathbb{R}^{D \times D}$,

$$\xi' = \boldsymbol{t}\, \mathrm{softmax}(\beta \boldsymbol{t}^{\top} \boldsymbol{G} \xi), \tag{7}$$

where β is the scale factor set to be sqrt(D). We assume the constraint that the query and memory should stay in the same space which is satisfied by $\mathbf{G}$ decomposed as $\mathbf{G} = \mathbf{M}^{\top}\mathbf{M}$. It can further be spectrally decomposed to get:

$$\xi' = \boldsymbol{t}\, \mathrm{softmax}(\beta \boldsymbol{t}^{\top} \boldsymbol{U}^{\top} \Lambda \boldsymbol{U} \xi) \tag{8}$$

where U is an orthogonal matrix and Λ is a positive diagonal matrix.

To achieve feature sensitive retrieval while factoring in all the information in the embeddings, we desire diag($\boldsymbol{\Lambda}$) not to be zero abounded, i.e., feature selection. To achieve that and also enable multi-modal retrieval, multiple $\boldsymbol{\Lambda}$s are learned and we desire $\prod_{i=1} \boldsymbol{\Lambda}_i$ close to the identity matrix such that if one feature is highlighted in one mode it should be suppressed in other modes. As such, the new attention mechanism becomes

$$\xi' = \boldsymbol{W} \underset{i=1}{\overset{s}{cat}} [\boldsymbol{t}\, \mathrm{softmax}(\beta \boldsymbol{t}^{\top} \boldsymbol{U}^{\top} \Lambda_i \boldsymbol{U} \xi)]. \tag{8}$$

For our model, ξ is the tokens from the master view, $\boldsymbol{t}$ is the tokens from reference view, s = 2, $\boldsymbol{W}$ projects tokens back to its original dimension and ξ' is the retrieved tokens from the reference view.

3.3 Learnable Epipolar Geometry

Token separability may be limited by the memory size and image content (e.g., existence of repetitive or uniform texture in the image). To further ensure secure retrieval without corrupting the encoded information, we constrain the attention mechanism with epipolar geometry through *gated positional cross attention* (GPCA) following [10]. In GPCA, positional embedding is modeled as trainable quadratic polynomial of relative positional encoding $\boldsymbol{v}_{pos}{}^{\top}\boldsymbol{r}_{ij}$. For regular stereo that is rectified, candidate retrievals reside within collinear horizontal lines. Therefore, we set $\boldsymbol{v}_{pos} = -\alpha(0,0,0,0,0,1,...,0)$, $\boldsymbol{r} = (1,\delta_1,\delta_2,\delta_1\delta_2,\delta_1^2,\delta_2^2,0,...,0)$, $W_Q = W_K = 0$, $W_V = I$.

In the equations above, $\boldsymbol{r}$ is the position vector of (δ_1, δ_2) which are the relative coordinates with respect to the query.

The locality strength $\alpha > 0$ determines how focused attention is along the horizontal line $\delta_2 = 0$. The positional attention scores are calculated as softmax normalized L_2 distance between the attended tokens and the query:

$$\boldsymbol{A}_{pos,ij} = \mathrm{softmax}(\boldsymbol{v}_{pos}^{\top}\boldsymbol{r}_{ij}). \tag{9}$$

With the learnable gating parameter λ, the GPCA attention scores are calculated as:

$$\boldsymbol{A}_{ij} = \mathrm{norm}[(1-\sigma(\lambda))\boldsymbol{A}_{cnt,ij} + \sigma(\lambda)\boldsymbol{A}_{pos,ij}], \tag{10}$$

$$\mathrm{norm}[x] = x_{ij} \Big/ \sum_k x_{ik} \tag{11}$$

where σ is the sigmoid function, and $\boldsymbol{A}_{cnt,ij}$ is the content attention score calculated by polarized attention. To avoid GPCA from being stuck at $\lambda >> 1$, we initialize $\lambda = 1$ for all layers.

3.4 Regularization

Matrix U has to be orthogonal to guarantee that query and memory are attended in the same space. However, U in each layer is trainable parameter; even though it can be initialized with orthogonal matrices, during training process the orthogonality may not hold. Therefore, we introduce an orthogonality regularization loss to U as:

$$L_o(\mathbf{U}) = \frac{1}{d^2} \,||\mathbf{U}^\top \mathbf{U} - \mathbf{I}||_F \tag{12}$$

where d is the size of U and $\|\cdot\|_F$ is the Frobenius norm of matrix. Although U can be orthogonalized through Cayley's parameterization, it is computational expensive for large matrix as inversion is involved and we found it is more difficult to converge and unstable in our case.

To induce the diagonal matrix $\boldsymbol{\Lambda}$ to be trained into the desired form, we modified Hoyer regularizer [29] to mitigate the proportional scaling issue and at the same time to pull $\boldsymbol{\Lambda}$ away from being identity matrix. We introduce the following regularization:

$$L_\Lambda(\Lambda) = \frac{\left| \prod_{i=1}^{s} |\Lambda_i|_e - \mathbf{I} \right|_1}{\prod_{i=1}^{s} ||\Lambda_i||_F}, \tag{13}$$

where $|\cdot|_e$ is the element-wise absolute function.

3.5 Training

In this section, we provide details of the training method we used. We closely followed the self-supervised stereo training method provided in [24]. The model is trained to predict the target image from the other viewpoint in a stereo pair. Unlike classical binocular and multi-view stereo methods, the image synthesis process in our case is constrained by predicted depth instead of disparity as an intermediary variable. Specifically, given a target image I_t, a source image $I_{t'}$, and the predicted depths D_t, through the relative pose between two views $T_{t \to t'}$ calculated with the provided stereo base width (0.54m for KITTI) and calibration information, the correspondent coordinates between two images can be calculated. Following [32], the target image can be reconstructed from source image using bilinear sampling, which is sub-differentiable.

The depth prediction should minimize the photometric reprojection error constructed for both master and reference view as follows:

$$L_p = \omega_m pe(I_t, I_{t' \to t}) + \omega_r pe(I_{t'}, I_{t \to t'}), \tag{14}$$

where ω_m (ω_r) is the weight for master (reference) view, and pe($\cdot$) is the photometric reconstruction error [63]:

$$\mathrm{pe}(\mathbf{X}, \mathbf{Y}) = \frac{\kappa}{2}(1 - \mathrm{SSIM}(\mathbf{X}, \mathbf{Y})) + (1 - \kappa) \, ||\mathbf{X} - \boldsymbol{Y}||_1 \tag{15}$$

κ= 0.85 and $I_{t' \to t}$ is the reprojected image:

$$\boldsymbol{I}_{t' \to t} = \mathrm{bi\text{-}sample} \langle \mathrm{proj}(\boldsymbol{D}_t, \boldsymbol{T}_{t \to t'}, \boldsymbol{K}) \rangle, \tag{16}$$

where **K** is the pre-computed intrinsic matrix, proj($\cdot$) is the resulting image coordinates projected from source view through

$$p'_t := \boldsymbol{K}\mathrm{T}_{t\to t'}\,\boldsymbol{D}_t[p_t]\boldsymbol{K}^{-1}p_t \tag{17}$$

and bi-sample⟨·⟩ is the bilinear sampler.

We also enforce edge-aware smoothness in the depths to improve depth-feature consistency defined as where $d_t^* = d/\mathrm{mean}(d_t)$ is the mean-normalized inverse depth in [60].

Unlike existing self-supervised stereo-matching methods that rely on predicted values to generate confidence map to detect occlusions, e.g., left-right consistency check, ChiTransformer detects occluded area on the fly in the form of heat map in rectification stage. During training, heat map from the last GPCA layer is up-sampled to the output resolution and used as a mask m_h in loss computation. For stereo training, static camera and synchronous movement between objects and camera are not issues, hence we do not apply the binary auto-masking to block out the static area in the image.

During inference, only the master ViT output is up-scaled and refined to make the prediction. While in the training stage, both ViT towers in ChiTransformer are trained in tandem to predict depth and calculate their own losses L_p and L_s.

Final Training Loss

By combining the reconstruction loss, per-pixel smoothness from two views and the regularizations for the matrices **U** and **Λ**, the final training loss is:

$$\mathscr{L} = \mathrm{mean}(m_h \odot L_p) + \mu_s L_s + \mu_o L_o + \mu_\lambda L_\lambda, \tag{18}$$

where μ_* are the hyperparameters that balance the contributions from different loss terms.

Our models are implemented in PyTorch. With pretrained resnet-50 patch feature extractor and partial refinement layers from [50], the model is trained for 30 epochs using Adam [36] with a batch size of 12 and input resolution of 640 ×192. We use learning rate 1e-5 for the ResNet-50 and 1e-4 for the rest part of the network in the first 20 epoch and then decayed to 1e-5 for the remaining epochs. We set $\omega_m = 0.6$, $\omega_r = 0.4$, μ_s=1e-4, μ_o=1e-7 and μ_Λ =1e-3.

4 EXPERIMENTS

The model is trained on KITTI 2015 [22]. We show that our model significantly improves accuracy compared to its top CNN-based counterparts. Side-by-side comparisons are given in this section with the state-of-the-art self-supervised stereo methods [41, 61, 62]. Ablation study is conducted to validate that several features in ChiTransformer contribute to the improved prediction. Finally, we extend our model to fisheye images and yield visually satisfactory result as shown in figure 7.

Table 1. Quantitative Results

	Method	NOC			ALL		
		D1 (bg)	D1 (fg)	D1 (all)	D1 (bg)	D1 (fg)	D1 (all)
Supervised	DispNet [46]	4.11	3.72	4.05	4.32	4.41	4.43
	GC-Net [34]	2.02	5.58	2.61	2.21	6.16	2.87
	iResNet [43]	2.07	2.76	2.19	2.25	3.40	2.44
	PSMNet [9]	1.71	4.31	2.14	1.86	4.62	2.32
Self-supervised	Yu et al. [33]	-	-	8.35	-	-	19.14
	Zhou et al. [74]	-	-	8.61	-	-	9.91
	SegStereo [65]	-	-	7.70	-	-	8.79
	OASM [41]	5.44	17.30	7.39	6.89	19.42	8.98
	PASMnet˙192 [62]	5.02	15.16	6.69	5.41	16.36	7.23
	Flow2Stere [45]	4.77	14.03	6.29	5.01	14.62	6.61
	pSGM [40]	4.20	10.08	5.17	4.84	11.64	5.97
	MC-CNN-WS [57]	3.06	9.42	4.11	3.78	10.93	4.97
	SsSMnet [73]	2.46	6.13	3.06	2.70	6.92	3.40
	PVSstereo [61]	**2.09**	5.73	2.69	**2.29**	6.50	2.99
	ChiT-8 (ours)	2.24	4.33	2.56	2.50	5.49	3.03
	ChiT-12 (ours)	2.11	**3.79**	**2.38**	2.34	**4.05**	**2.60**

Comparison of our model to the state-of-the-art self-supervised binocular stereo methods. Lower is better for all metrics.

4.1 KITTI 2015 Eigen Split

We divide the KITTI dataset following the method of Eigen et al. [18]. Same intrinsic parameters are applied to all images by setting the camera principal point at the image center and the focal length as the average focal length of KITTI. For stereo training, the relative pose of a stereopair is set to be pure horizontal translation of a fixed length (0.54m) according to the KITTI sensor setup. For a fair comparison, depth is truncated to 80m according to standard practice [23]

4.2 Quantitative Results

We compare the results of the two different configurations of our model with state-of-the-art self-supervised stereo approaches. ChiT-8 has 4 SA layers followed by 4 rectification blocks, while ChiT-12 has 6 SA layers and 6 rectification blocks. The results in Table 1 show that ChiTransformer outperforms most of the existing methods, particularly in the prediction of the foreground regions. This trait is as expected that foreground regions are more likely to be abounded by distinctive features that benefit depth cue rectification. We show qualitative result in Figure 5-6. With content information from both views, ChiTransformer provides more details consistent to the image features compared to existing self-supervised methods.

We also compare our method with top self-stereo-supervised MDE methods to show the reliability gain in terms of accuracy improvement. For a fair comparison, we choose the models that are trained on KITTI with stereo supervision. Methods trained over multiple datasets are not considered. Quantitative results are shown in Table 2. Side-by-side prediction comparison is shown in Figure 2-4.

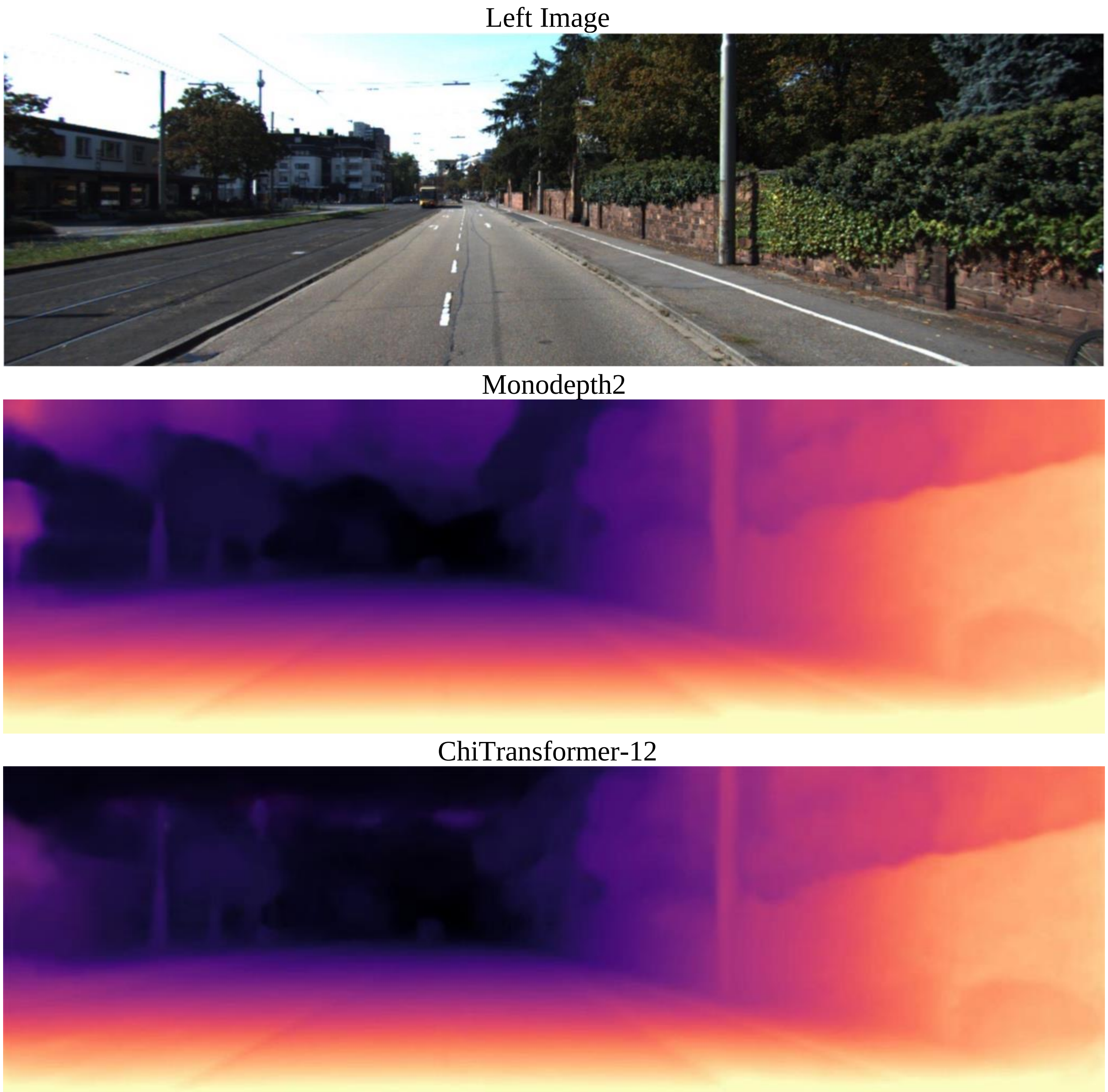


Figure 3. Sample results compared with self-stereo-supervised fully-convolutional network Monodepth2. ChiTransformer shows better global coherence (e.g., sky region, sides of image) and provides feature consistent details.

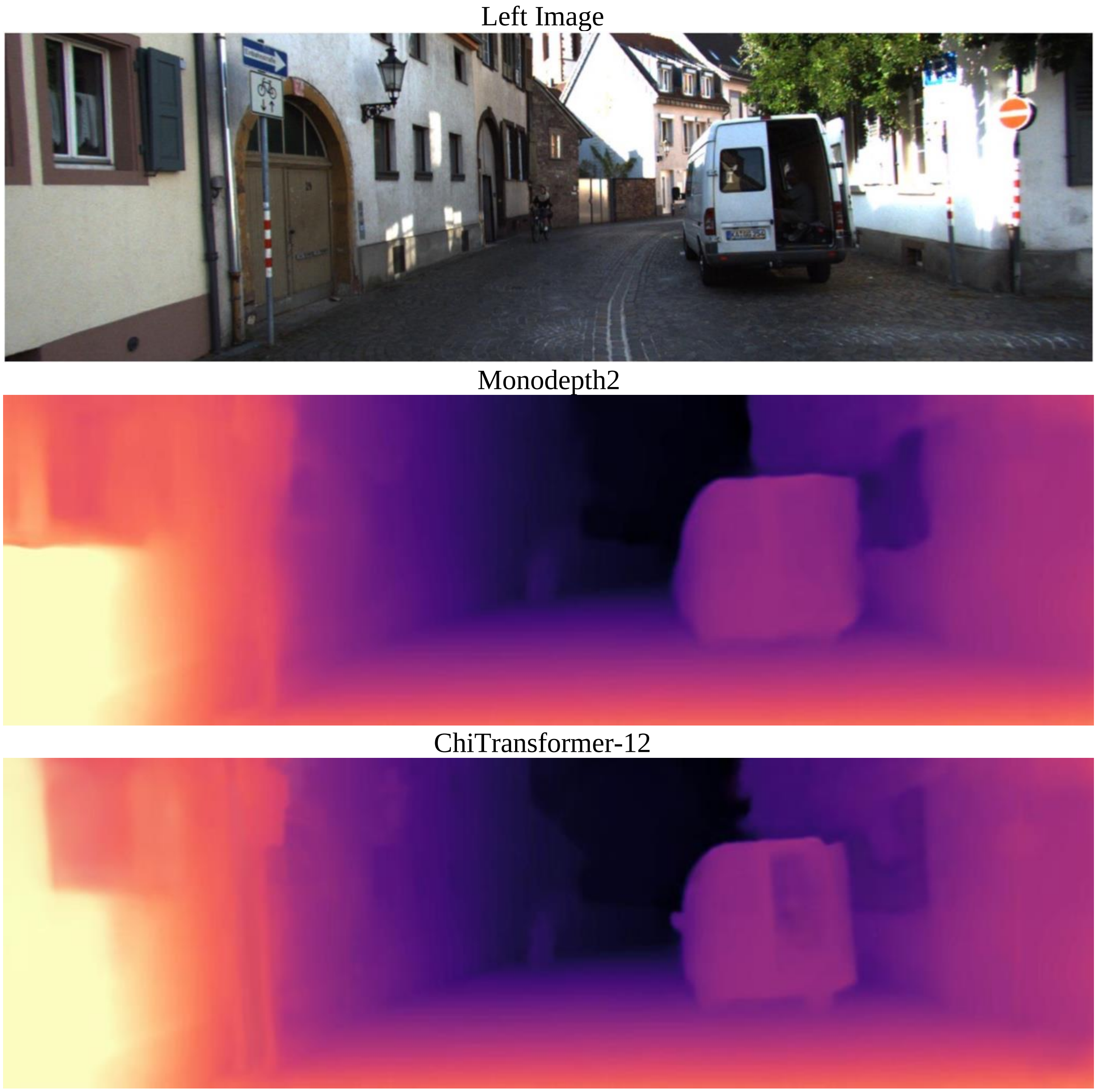


Figure 4. Sample results compared with self-stereo-supervised fully-convolutional network Monodepth2. ChiTransformer shows better global coherence (e.g., sky region, sides of image) and provides feature consistent details.

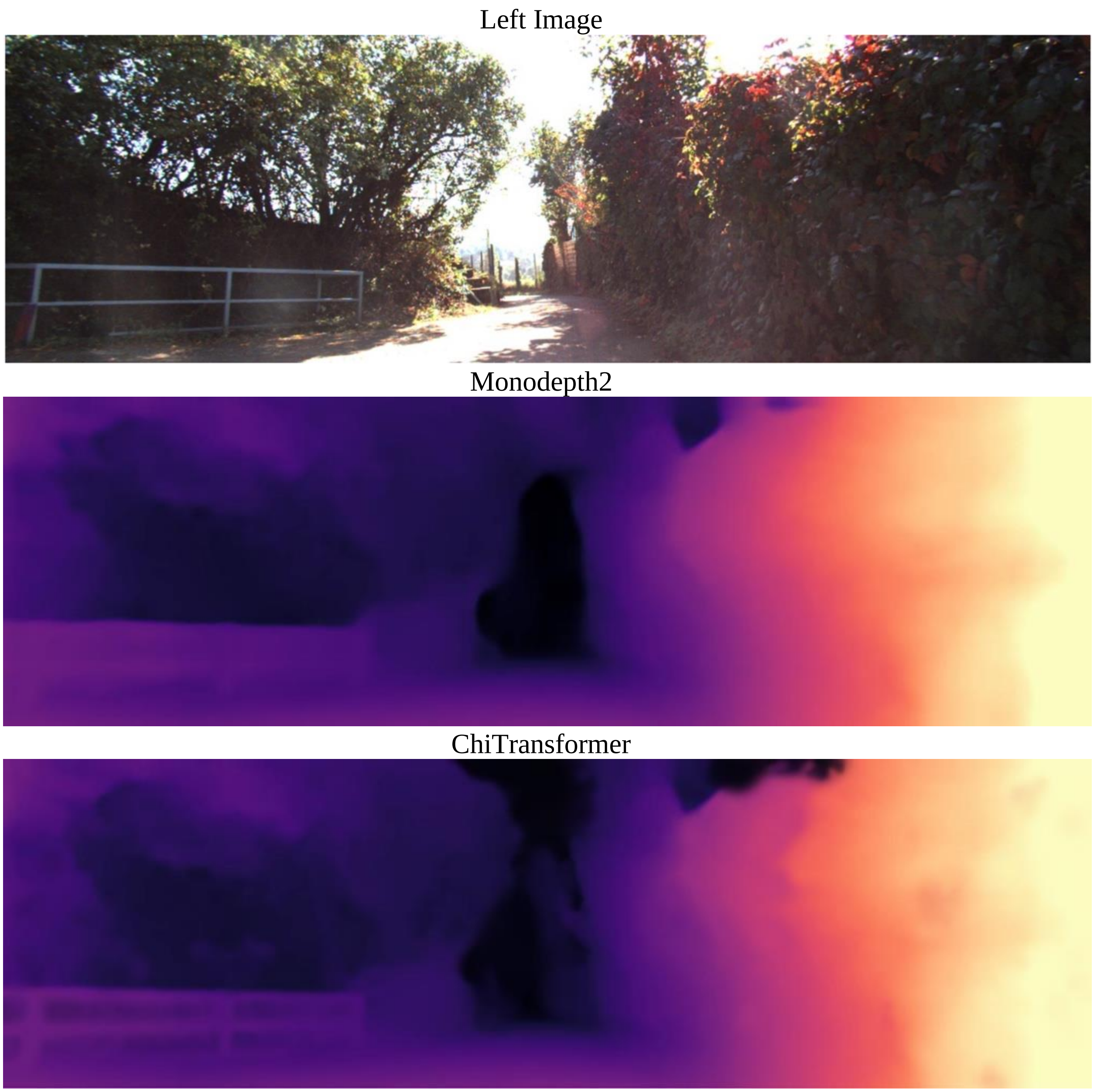


Figure 5. Sample results compared with self-stereo-supervised fully-convolutional network Monodepth2. ChiTransformer shows better global coherence (e.g., sky region, sides of image) and provides feature consistent details

Table 2. Comparison with Stereo Self-supervised Monocular Methods

Method	AbsRel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Garg et al. [21]	0.152	1.226	5.849	0.246	0.784	0.921	0.967
3Net(R50) [48]	0.129	0.996	5.281	0.223	0.831	0.939	0.974
3Net(VGG) [48]	0.119	1.201	5.888	0.208	0.844	0.941	0.978
SuperDepth+pp (1024×382) [47]	0.112	0.875	4.968	0.207	0.852	0.947	0.977
Monodepth2 [24]	0.109	0.873	4.960	0.209	0.864	0.948	0.975
ChiT-12 (ours)	**0.073**	**0.634**	**3.105**	**0.118**	**0.924**	**0.989**	**0.997**

All models listed in the table above are trained with self-supervised methods using stereo pair. Same as the monocular methods, ChiTransformer relies on depth cues to estimate depth, only with extra information from a second image.

4.3 Ablation Study

To understand how each major feature contributes to the overall performance of ChiTransformer, ablation study is conducted by suppressing or activating specific components of the model. We observe that each component in our model is designed to push the performance a bit forward which aggregates into a sizable improvement. Here we provide some insights on the major features based on observation.

4.3.1 Self-attention layer

largely improves the separability of each token with long range complex contextual information. Without SA layer, the retrieval process would take up a hopping behavior and render erroneous predictions.

4.3.2 Polarized attention

We learn the matrix **G** through its spectral decomposition to gain more control over its behavior. Direct learning of **G** tends to result in feature negligence as the major features contained within a token dominate or take all the reward. With the complementary feature highlighting-suppressing strategy as we desire $\prod\mathbf{\Lambda}$ to be close to $\mathbf{I}$, the features from both parties can be attended. Meanwhile, since $\mathbf{\Lambda}$ is not porous with zeros, i.e., no Lasso regularization involved, all information contained in the token is more or less attended.

4.3.3 Learnable Epipolar Geometry

Intuitively, for rectified stereo a pixel pair is guaranteed to reside in the same horizontal line and hence the attending space should be that line. However, the slotted attention region hurts the inter-line connection and cause serrated effects on vertical features even that feature is distinctive, e.g., an edge. Whilst the learnable epipolar geometry in GPCA solves the problem by allowing global but focused view over the lines and at the same time further improves the inter-line separability. Quantitative results are given in Table 3.

Table 3. Ablation Study

	AbsRel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
ChiT+P	0.106	4.845	0.204	0.878	0.960	0.981
ChiT+G+LEG	0.101	4.783	0.203	0.895	0.966	0.983
ChiT+P+Linear	0.092	4.535	0.201	0.889	0.964	0.987
ChiT+P+LEG	0.085	3.924	0.181	0.906	0.979	0.991

Evaluations for different settings of ChiTransformer (ChiT) trained on KITTI 2015 with Eigen split. "P" denotes the polarized attention. "G" stands for the direct learning of matrix G. "LEG" represents the feature of learnable epipolar geometry. "Linear" is the single line attention zone. ChiT with only P enabled has the lowest score due to inferior token separability. With the addition of LEG, the model: ChiT+P+LEG, becomes the top performer and show the advantage of P over G compared with ChiT+G+LEG. ChiT+P+Linear has the 2nd best performance. The serration effect due to "Linear" is largely mitigated by the long-range context information and SA layers.

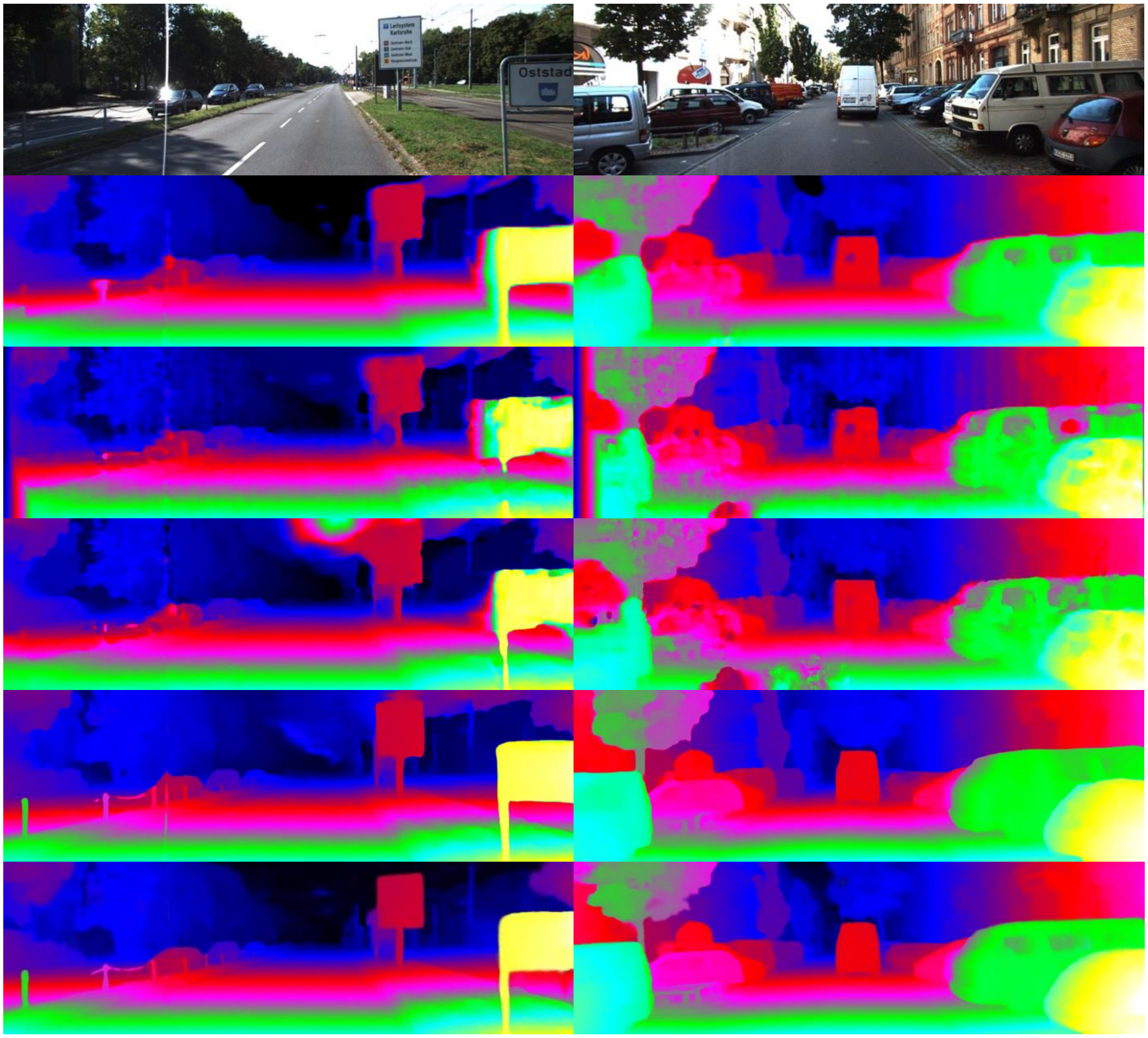


Figure 6. Qualitative results on the KITTI stereo 2015 compared with existing self-supervised stereo matching methods. The sharp depth maps generated by our model (ChiTransformer-12 in the second last row) provides more reliable estimation especially in the close range as reflected in the error map and better global coherence consistent.

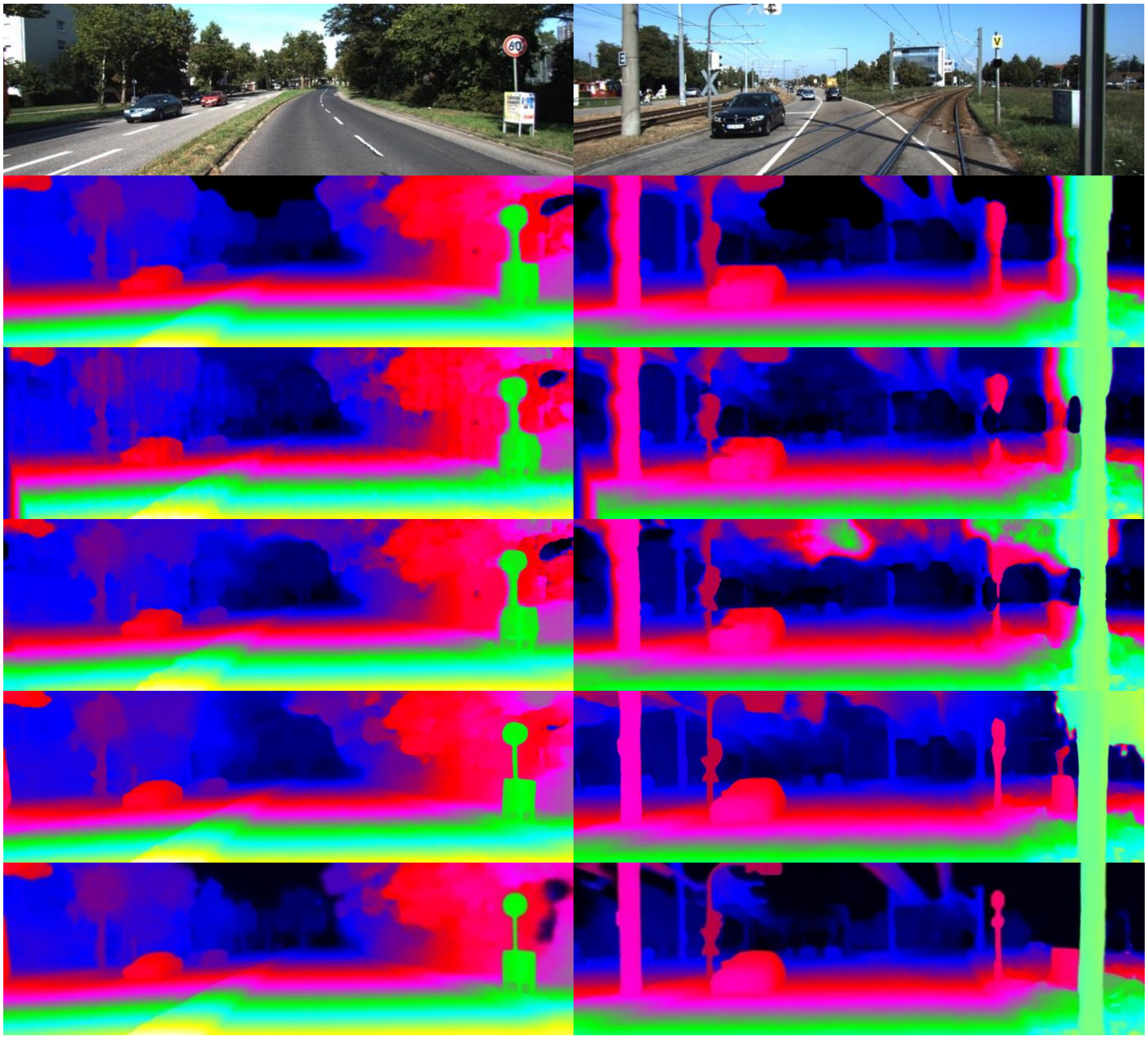
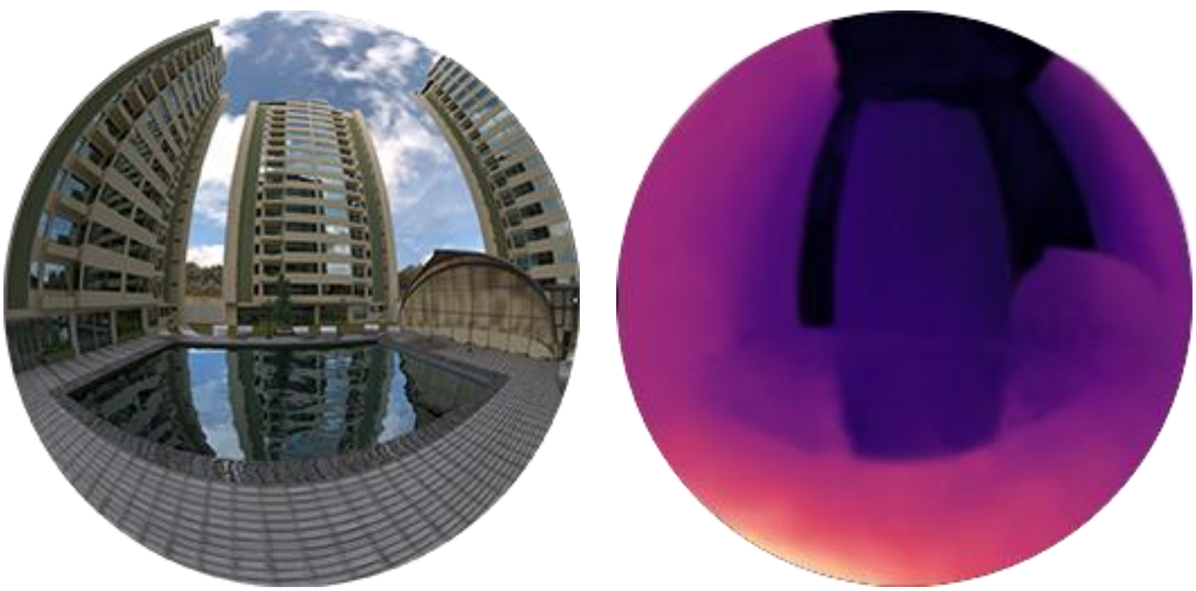


Figure 7. Qualitative results on the KITTI stereo 2015 compared with existing self-supervised stereo matching methods. The sharp depth maps generated by our model (ChiTransformer-12 in the second last row) provides more reliable estimation especially in the close range as reflected in the error map and better global coherence consistent.

5 CONCLUSION

By investigating the limitations of the two prevalent methodologies of depth estimation, we present ChiTransformer, a novel and versatile stereo model that generates reliable depth estimation with rectified depth cues instead of stereo matching. With the three major contributions: (1) polarized attention mechanism, (2) learnable epipolar geometry, and (3) the depth cue rectification method, our model outperforms the existing self-supervised stereo methods and achieves state-of-the-art accuracy. In addition, due to its versatility, ChiTransformer can be applied to fisheye images without warping, yielding visually satisfactory results.



Synthetic fisheye image — Planar depth estimation

Figure 8. Example result of ChiTransformer for fisheye depth estimation. With learnable epipolar curve $v_{pos,ij} = (1, a, b, c, d, e)_{ij}$ (constant term is set to 1 to avoid proportional scaling) and circular masks, ChiTransformer can directly work on circular image without warping.

REFERENCES

References

1. G. Goh A. Ramesh, M. Pavlov and tech. rep. S. Gray. Dall·e: Creating images from text. In OpenAI, 2021.
2. Aria Ahmadi and Ioannis Patras. Unsupervised convolutional neural networks for motion estimation. In2016 IEEE international conference on image processing (ICIP), pages 1629–1633. IEEE, 2016.
3. Alex M Andrew. Multiple view geometry in computer vision. Kybernetes, 2001.
4. Stephen T Barnard and Martin A Fischler. Computational stereo. ACM Computing Surveys (CSUR), 14(4):553–572, 1982.
5. Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4009–4018, 2021.
6. Amlaan Bhoi. Monocular depth estimation: A survey. arXiv preprint arXiv:1901.09402, 2019.
7. Arunkumar Byravan and Dieter Fox. Se3-nets: Learning rigid body motion using deep neural networks. In2017 IEEE International Conference on Robotics and Automation (ICRA), pages 173–180. IEEE, 2017.
8. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. InEuropean Conference on Computer Vision, pages 213–229. Springer, 2020.

9. 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.

10. Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. arXiv preprint arXiv:1911.03584, 2019.

11. St´ephane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision trans-formers with soft convolutional inductive biases. arXiv preprint arXiv:2103.10697, 2021.

12. Mete Demircigil, Judith Heusel, Matthias L¨owe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. Journal of Statistical Physics, 168(2):288–299, 2017.

13. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

14. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition atscale.arXiv preprint arXiv:2010.11929, 2020. 2, 3

15. Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 2758–2766, 2015

16. Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deep pruner: Learning efficient stereo matching via differentiable patch match. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4384–4393, 2019

17. Andrea Eichenseer and Andr´e Kaup. A data set providing synthetic and real-world fisheye video sequences. In2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1541–1545. IEEE, 2016. 2

18. David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. arXiv preprint arXiv:1406.2283, 2014.

19. Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip H¨ausser, Caner Hazırbas ̧, Vladimir Golkov, Patrick Van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. arXiv preprint arXiv:1504.06852, 2015.

20. Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2002–2011, 2018

21. Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In European conference on computer vision, pages 740–756. Springer, 2016

22. Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In2012IEEE conference on computer vision and pattern recognition, pages 3354–3361. IEEE, 2012

23. Cl´ement Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 270–279, 2017. 2, 3, 8

24. Cl´ement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3828–3838, 2019.

25. Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 484–500, 2018.

26. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

27. Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.

28. Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2651–2660, 2019.

29. Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. Journal of machine learning research, 5(9), 2004

30. Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2821–2830, 2018.

31. Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In Proceedings of the European Conference on Computer Vision (ECCV), pages614–630, 2018.

32. Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. Advances in neural information processing systems, 28:2017–2025, 2015.

33. J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In European Conference on Computer Vision, pages 3–10. Springer, 2016.

34. 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.

35. Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun. A guide to convolutional neural networks for computer vision. Synthesis Lectures on Computer Vision, 8(1):1–207, 2018.

36. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

37. Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. Advances in neural information processing systems, 29:1172–1180, 2016.

38. 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, 2020.

39. Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326, 2019.

40. Yeongmin Lee, Min-Gyu Park, Youngbae Hwang, Youngsoo Shin, and Chong-Min Kyung. Memory-efficient parametric semiglobal matching. IEEE Signal Processing Letters, 25(2):194–198, 2017.

41. Ang Li and Zejian Yuan. Occlusion aware stereo matching via cooperative unsupervised learning. In Asian Conference on Computer Vision, pages 197–213. Springer, 2018.

42. Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang. Deep attention-based classification network for robust depth prediction. In Asian Conference on Computer Vision, pages 663–678. Springer, 2018. 1

43. 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, pages2811–2820, 2018.

44. Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1925–1934, 2017.

45. Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6648–6657, 2020.

46. 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.

47. Sudeep Pillai, Rareș Ambruș, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In2019International Conference on Robotics and Automation (ICRA), pages 9250–9256. IEEE, 2019.

48. Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In 2018 International conference on 3d vision (3DV), pages 324–333. IEEE, 2018.

49. Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. arXiv preprint arXiv:2008.02217, 2020.

50. René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 12179–12188, 2021.

51. René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv preprint arXiv:1907.01341, 2019.

52. Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In Advances in neural information processing systems, pages 14866–14876, 2019. 2

53. Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision, 47(1):7–42, 2002.

54. Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In European conference on computer vision, pages 746–760. Springer, 2012.

55. Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In2012 IEEE/RSJ international conference on intelligent robots and systems, pages 573–580. IEEE, 2012.

56. Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR,2021.

57. Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. Weakly supervised learning of deep metrics for stereo reconstruction. In Proceedings of the IEEE International Conference on Computer Vision, pages 1339–1348, 2017.

58. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

59. Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. arXiv preprint arXiv:1704.07804, 2017

60. Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2022–2030, 2018.

61. Hengli Wang, Rui Fan, Peide Cai, and Ming Liu. Pvstereo: Pyramid voting module for end-to-end self-supervised stereo matching.IEEE Robotics and Automation Letters, 6(3):4353–4360, 2021.

62. Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. IEEE transactions on pattern analysis and machine intelligence,

63. Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004.

64. Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5791–5800, 2020.

65. Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 636–651, 2018.

66. Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV), pages 767–783, 2018.

67. Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10502–10511, 2019.

68. Wei Yin, Yifan Liu, Chun hua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5684–5693, 2019. 1

69. Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1983–1992, 2018.

70. 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.

71. Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In European Conference on Computer Vision, pages 528–543. Springer, 2020.

72. 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/CVF Conference on Computer Vision and Pattern Recognition, pages 185–194, 2019.

73. Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability.arXiv preprintarXiv:1709.00930, 2017.

74. Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In Proceedings of the IEEE International Conference on Computer Vision, pages 1567–1575, 2017

75. Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.