DualNet: Robust Self-Supervised Stereo Matching with Pseudo-Label Supervision

Yun Wang[1], Jiahao Zheng[1], Chenghao Zhang[2], Zhanjie Zhang[3], Kunhong Li[4], Yongjian Zhang[4] Junjie Hu[5,6*]

[1] City University of Hong Kong
[2] Chinese Academy of Sciences Institute of Automation, CASIA
[3] Zhejiang University
[4] Sun Yat-sen University
[5] Shenzhen Institute of Artificial Intelligence and Robotics for Society
[6] The Chinese University of Hong Kong (Shenzhen)
{ywang3875-c, jhzheng4-c}@my.cityu.edu.hk, zhangchenghao18@mails.ucas.ac.cn, cszzj@zju.edu.cn,
{likh25,zhangyj85}@mail2.sysu.edu.cn, hujunjie@cuhk.edu.cn

Abstract

Self-supervised stereo matching has drawn attention due to its ability to estimate disparity without needing ground-truth data. However, existing self-supervised stereo matching methods heavily rely on the photo-metric consistency assumption, which is vulnerable to natural disturbances, resulting in ambiguous supervision and inferior performance compared to the supervised ones. To relax the limitation of the photo-metric consistency assumption and even bypass this assumption, we propose a novel self-supervised framework named DualNet, which consists of two key steps: robust self-supervised teacher learning and pseudo-label supervised student training. Specifically, the teacher model is first trained in a self-supervised manner with a focus on feature-metric consistency and data augmentation consistency. Then, the output of the teacher model is geometrically constrained to obtain high-quality pseudo labels. Benefiting from these high-quality pseudo labels, the student model can outperform its teacher model by a large margin. With the two well-designed steps, the proposed framework DualNet ranks 1^{st} among all self-supervised methods on multiple benchmarks, surprisingly even outperforming several supervised counterparts.

Introduction

Stereo matching aims to find corresponding pixels between a rectified image pair, which is crucial for various applications, e.g., AR, robotics, and navigation. Recently, learning-based supervised stereo matching methods (Li et al. 2022; Wang et al. 2024a,b, 2025) have obtained remarkable results. However, these methods are highly dependent on large-scale ground-truth labels for supervision, the acquisition of which is a significant challenge (Yang et al. 2019; Zhang et al. 2024a, 2025). As an alternative, self-supervised learning approaches have emerged, offering competitive accuracy without the need for ground-truth disparity labels.

Existing self-supervised stereo matching methods (Zhong, Dai, and Li 2017; Wang et al. 2020; Su and Ji 2022) rely on the photo-metric consistency assumption, which assumes that the appearance of a point in 3D space is color-invariant across different views. The hypothesis of photo-metric consistency is to maximize the similarity between the left image and the reconstructed left image after being warped from the right perspective. However, this foundational premise can be compromised by various real-world factors, such as *low texture, occlusion, reflections, and illumination changes,* as illustrated in Fig. 2. These challenges can confuse the self-supervision loss, leading to ambiguous supervision and ultimately resulting in inferior performance compared to supervised methods.

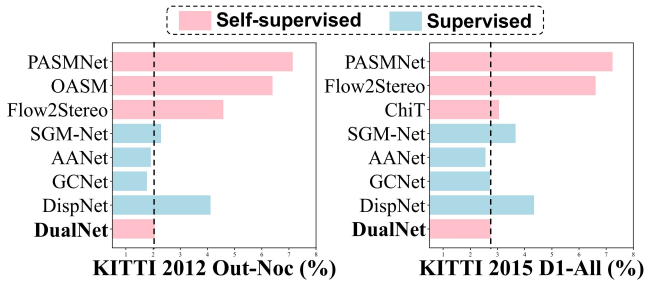

Figure 1: Visualized performance comparisons of state-of-the-art stereo matching methods on KITTI 2012 and KITTI 2015 benchmarks (**lower is better**).

In this work, we address the challenges of self-supervised stereo matching by introducing novel learning strategies. We argue that stereo matching inherently requires establishing dense correspondences along epipolar lines. The features derived from this matching process allow for the encoding of *textureless* patterns with sufficient local discriminability. Consequently, these learned features, which capture distinctive textures, can serve as effective supervisory signals. To this end, we propose a feature-metric consistency loss that enforces consistency between the features of the left view and its reconstructed counterpart, analogous to the photo-metric consistency loss.

To further address ambiguities arising from problematic regions, such as *occlusions* and *illumination changes*, we explore the application of contrastive learning, which has demonstrated success in self-supervised learning. However, the integration of contrastive learning into self-

*Corresponding author is Junjie Hu.

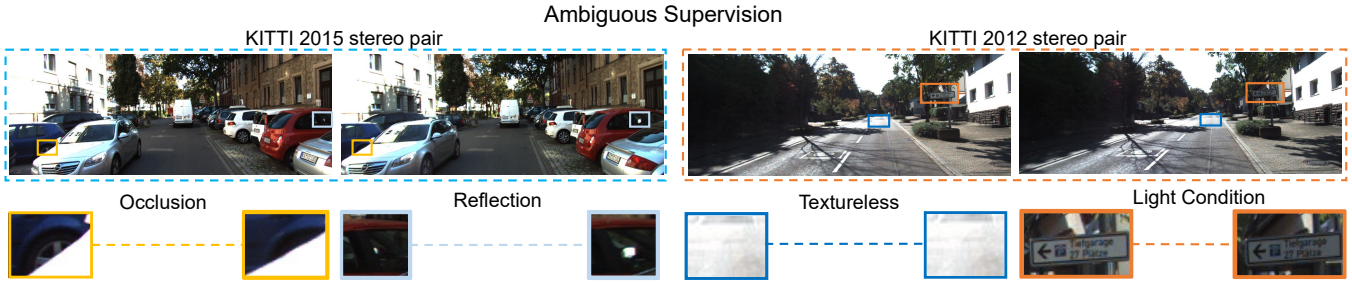

Figure 2: Illustration of the ambiguity supervision problem in self-supervised stereo matching on the KITTI stereo dataset. At the bottom of each image, we mark with color rectangles and zoom into these problematic areas for better visualization.

supervised stereo matching remains under-explored. Recent studies (Afham et al. 2022) suggest that hard positive samples can significantly enhance contrastive learning. Motivated by this insight, we introduce a data augmentation branch that generates hard positive samples by incorporating variations in illumination and asymmetric occlusions, compared to the regular image pairs. We then propose a data augmentation loss to enforce consistency between the outputs of the augmented branch and the regular branch. The underlying intuition is that the augmented image pairs, despite their challenges, share the same contextual information as the regular pairs and should produce consistent disparities. By integrating the feature-metric consistency and the data augmentation consistency losses into the self-supervised stereo matching framework, we aim to develop a network that is more robust and context-aware, rather than relying solely on photo-metric consistency in idealized regions.

Furthermore, we propose a pseudo-label supervision mechanism to bypass the ambiguous supervision problem in photo-metric consistency by exploiting high-quality pseudo labels. A key observation behind the proposed mechanism is that deep stereo models have empirically demonstrated success in training even by deploying only sparse ground truth labels or even sparse noisy estimates (Shen et al. 2023). For example, deep models can achieve impressive results on KITTI datasets with sparse Ground Truth (less than 30% is annotated for 200 training images). Besides, since stereo matching can be viewed as a disparity probability prediction problem (Zhang et al. 2019), we convert the disparity maps derived from the robust self-supervised teacher model into a pseudo disparity probability distribution, modeled as a uni-modal distribution. Then, the disparity distribution of the student model is used to align with that of the teacher model. As a result, the student model can surpass its teacher and even outperform some supervised methods.

Overall, the proposed learning strategies can be divided into two key steps: (1) robust self-supervised teacher learning, which incorporates feature-metric and data augmentation consistency loss, and (2) pseudo-label supervised student training. Accordingly, we refer to our approach as ***DualNet***. As a result, our ***DualNet*** achieves 1^{st} performance among all self-supervised methods on KITTI 2012, KITTI 2015, Middlebury, and ETH3D benchmarks, even outperforming several supervised methods, as shown in Fig. 1.

In summary, our main contributions are:

- We present a self-supervised pipeline named ***DualNet***, which can be seamlessly applied to most stereo models, achieving State-of-the-Art (SOTA) performance among all self-supervised methods on four benchmark datasets.
- We introduce a feature-metric consistency loss and data augmentation consistency loss to encourage the model to be robust and context-aware against common natural disturbances in photo-metric consistency.
- We propose a pseudo-label supervision scheme that aligns the probability distribution of the student model with that of the teacher model to improve performance.

Related Works

Supervised Stereo Matching

Recently, CFNet (Shen and Dai 2021) and PCWNet (Shen et al. 2022) propose a multi-scale cost volume strategy to progressively narrow the disparity search space to alleviate high computational consumption. Motivated by RAFT (Teed and Deng 2020), the very recent CREStereo (Li et al. 2022), IGEV-Stereo (Xu et al. 2023) and Selective-IGEV (Wang et al. 2024a) adopt cascaded recurrent network to update the disparity iteratively. However, all the above methods rely on dense annotation labels, which are not always available and are often expensive to acquire in real-world settings.

Self-supervised Stereo Matching

SssMnet (Zhong, Dai, and Li 2017) proposes a loop photo-metric consistency loss. Segstereo (Yang et al. 2018) proposes to incorporate semantic cues to guide stereo matching. Flow2Stereo (Liu et al. 2020) proposes a unified method to jointly learn optical flow and stereo matching. OASM (Li et al. 2021) proposes to use occlusion cues as a depth cue for stereo matching. PAM (Wang et al. 2020) proposes a parallax attention mechanism to capture the stereo correspondence without the limitation of disparity variations. ChiTransformer (Su and Ji 2022) uses monocular depth cues and vision transformer (ViT) with cross-attention (Zhang et al. 2024b) to enhance stereo matching. However, all these methods hold the photo-metric consistency assumption to guide the network. Multiple real-world factors can violate this hypothesis, i.e., *occlusion, reflections, low texture, and illumination changes.* As a result, the ideal self-supervision loss suffers from ambiguous supervision in complex scenarios.

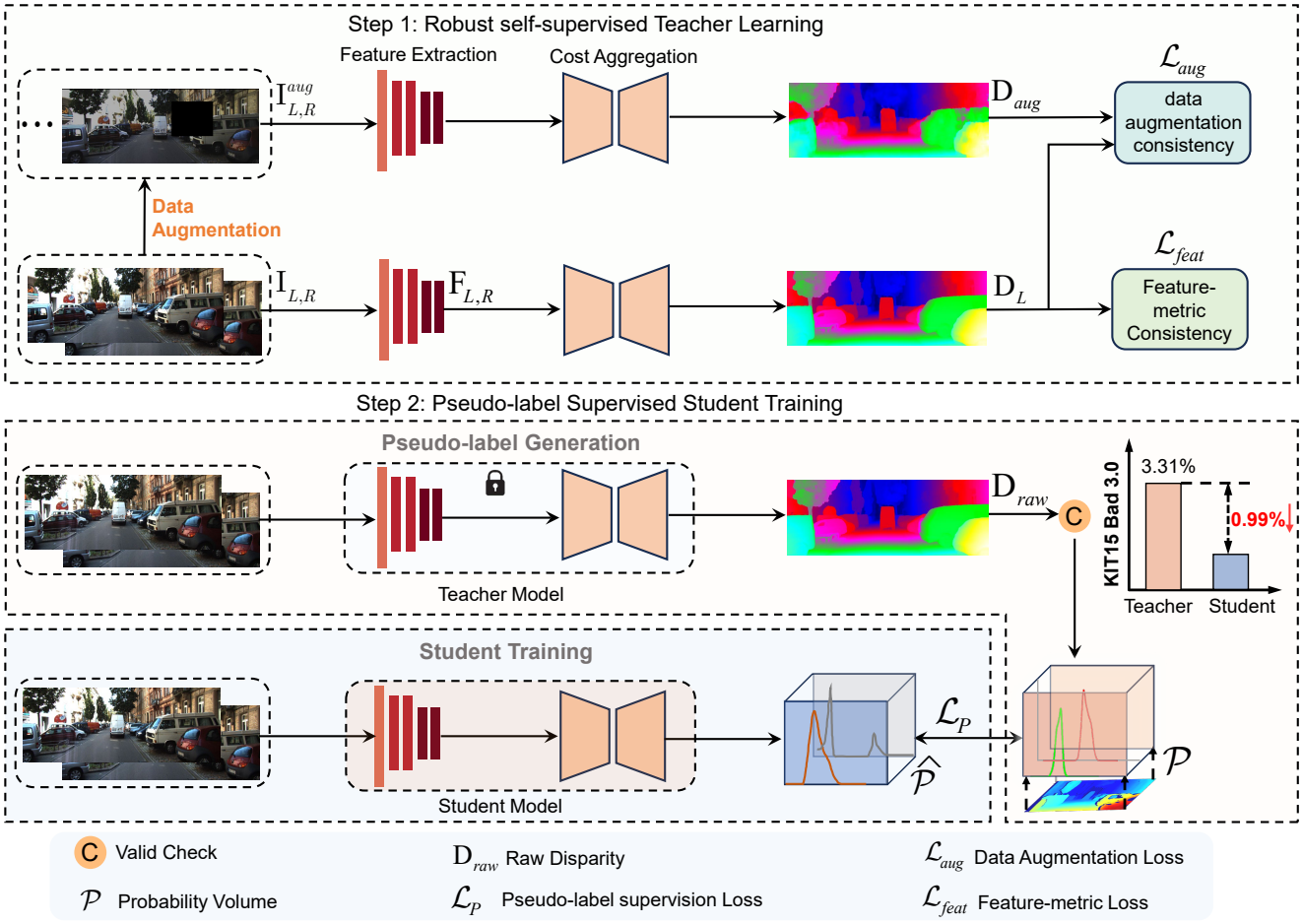

Figure 3: An overview of our DualNet. The first step is self-supervised teacher learning. The second step is pseudo-label supervised student training, which includes pseudo-label generation and student training.

Pseudo-label for Depth Estimation

The core idea of pseudo-labeling leverages the knowledge derived from labeled data to infer pseudo Ground Truth for unlabeled data, which can applied to self-supervised (Ding et al. 2022) and domain adaptation tasks (Shen et al. 2023; Yen et al. 2022). KD-MVS (Ding et al. 2022) devise a filtering scheme based on depth reprojection error in the context of multi-view stereo. UCFNet (Shen et al. 2023) uses an uncertainty estimation network to retain the high-confidence pixels of predicted disparity maps to align the domain gap. However, directly supervising the disparity maps will make the cost volume less constrained (Zhang et al. 2019). Our methodology uses high-quality pseudo labels to model a unimodal distribution, which directly supervises the disparity probability distribution of the student network.

Methodology

In this section, we first describe the proposed self-supervised pipeline named ***DualNet***, which mainly consists of two steps: 1) self-supervised teacher learning and 2) pseudo-label supervised student training. For step 1, we describe the proposed feature-metric and data augmentation consistency loss. For step 2, we depict the pseudo-label generation scheme and the proposed pseudo-label supervision loss.

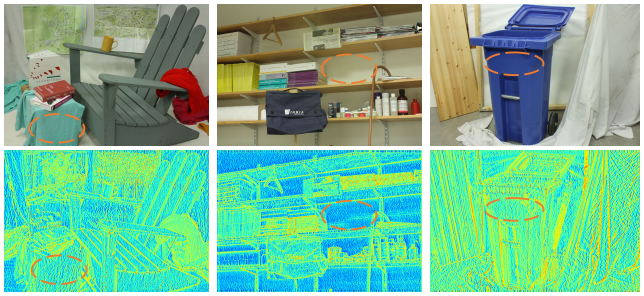

Figure 4: A visualization of learned visual representation. Dimension reduction of features is done by PCA.

Self-supervised Teacher Learning

We utilize CFNet (Shen and Dai 2021) as our backbone, and the overall architecture is shown in Fig. 3. For step 1, the framework first extracts features from regular image pairs $I_{L,R}$ and augmented image pairs $I^{aug}_{L,R}$, respectively. A cost

volume is then constructed by concatenating or correlating the left features, and the shifted right features along the disparity dimension. The resultant disparity map $\mathbf{D}_L$ and $\mathbf{D}_{aug}$ are obtained by performing cost aggregation. In addition to the baseline photo-metric consistency loss $\mathcal{L}_{photo}$ and the disparity smooth loss $\mathcal{L}_s$, we introduce the proposed feature-metric $\mathcal{L}_{feat}$ and data augmentation consistency loss $\mathcal{L}_{aug}$.

Feature-metric Consistency. Since the vanilla photo-metric consistency only contains RGB information, it is insufficient to distinguish *textureless* regions. A better solution is to implement the learned features (usually 128 dimensions) to encode textureless patterns to exhibit local discriminability. In Fig. 4, the dash-circled regions can be easily distinguished by the learned features (the bottom 3 pics). Therefore, the feature-metric consistency loss is introduced to formulate a more discriminative loss function.

Given the extracted features $\mathbf{F}_L$, $\mathbf{F}_R$ from the regular samples, the reconstructed features $\hat{\mathbf{F}}_L$ can be generated as:

$$\hat{\mathbf{F}}_L(i,j) = \mathbf{F}_R(i + \mathbf{D}_L(i,j), j), \tag{1}$$

where (*i, j*) represent the pixel coordinates and $\hat{\mathbf{F}}_L$ is the reconstructed left features. Moreover, to evaluate its pixel-wise estimation confidence, we use the disparity probability volume to obtain a probability map $\mathbf{P}_m$ by taking the probability sum over the top 4 disparity candidates regarded to disparity estimation (Bangunharcana et al. 2021). Then, we generate a binary confidence mask $\mathbf{M}_c$ as follows:

$$\mathbf{M}_c(p) = \begin{cases} 1, & \mathbf{P}_m(p) > \delta \\ 0, & otherwise \end{cases}, \tag{2}$$

where p denotes the pixel location and δ is set to 0.8 in our settings. The confidence mask $\mathbf{M}_c$ will be used to regularize the feature-metric loss and data augmentation loss. Consequently, the proposed feature-metric loss can be defined as:

$$\mathcal{L}_{feat} = \sum_{i,j} ||(\mathbf{F}_L(i,j) - \hat{\mathbf{F}}_L(i,j)) \odot \mathbf{M}_c(i,j)||, \tag{3}$$

where $\odot$ is an element-wise product.

Data Augmentation Consistency. With the proposed feature-metric loss, the stereo network mitigates the limitation of the photo-metric loss towards *texturelessness*. However, the model may still suffer from invalid supervision due to *reflections, occlusions, and illumination changes* leading to performance degradation. We notice that recent works (He et al. 2020) in contrastive learning demonstrate that bringing challenging samples in self-supervised learning can provide robustness towards variations. Therefore, we adopt data augmentation strategies in stereo matching to construct such challenging samples $\mathbf{I}^{aug}_{L,R}$ from regular samples $\mathbf{I}_{L,R}$. Then, we enforce the disparity consistency between the output of regular samples and that of challenging samples to improve the model's robustness and context awareness.

Briefly, the data augmentation strategies include random variations in color, brightness, luminance, and asymmetric blur. Moreover, we randomly generate binary small crop masks to mask some regions on the right view $\mathbf{I}^{aug}_R$ to generate the pseudo occluded regions. Similar to the core of contrastive learning, the data augmentation consistency is ensured by minimizing the difference between $\mathbf{D}_L$ and $\mathbf{D}_{aug}$:

$$\mathcal{L}_{aug} = \text{Smooth}_{\text{L1}}\left(\mathbf{D}_{aug} \odot \mathbf{M}_{\text{c}}, \mathbf{D}_L \odot \mathbf{M}_{\text{c}}\right), \tag{4}$$

The final loss function for this step is defined as follows:

$$\mathcal{L}_{self} = \lambda_1 \mathcal{L}_{photo} + \lambda_2 \mathcal{L}_s + \lambda_3 \mathcal{L}_{feat} + \lambda_4 \mathcal{L}_{aug}, \tag{5}$$

where we set λ_1 = λ_3 = λ_4=1, and λ_2 = 10.

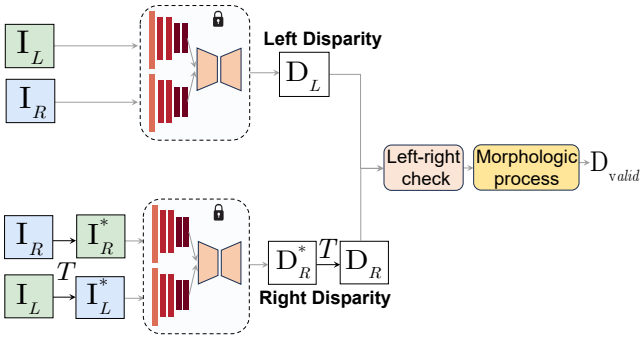

Figure 5: An pipeline of valid check. T denotes the horizontal flip operation.

Pseudo-label Supervised Student Training

In step 1, we incorporate the photo-metric consistency assumption, which can lead to performance degradation in challenging areas. To this end, we address the potential adverse effects by developing a valid supervisory signal to work around this assumption. Specifically, we incorporate the idea of pseudo-label supervision and use the pseudo probability distribution of the teacher model to align with that of the student model. Our intuition is that deep stereo methods have been empirically demonstrated to be successful in training using only sparse Ground Truth labels or even sparse noisy estimates (Aleotti et al. 2020; Tonioni et al. 2019). Therefore, the two terms of pseudo-label supervision, i.e., pseudo-label generation and student training, are employed to train the final model.

Pseudo-label Generation. Since the generated disparity map may have outliers, we employ a filtering procedure to filter outliers, as shown in Fig. 5. Specifically, we apply the horizontal flip that inverted original image pairs fed into the same model to estimate the right disparity map $\mathbf{D}_R$. Considering an arbitrary point $p(i,j)$ has d_1 in the left disparity map $\mathbf{D}_L$ and the disparity value of the corresponding point $p(i-d_1, j)$ in the right disparity map $\mathbf{D}_R$ is denoted by d_2. Thus, the warp error at $p(i,j)$ can be written as $e_{warp} = ||d_1 - d_2||$ and the filtered disparity map is formulated as follows:

$$\mathbf{D}_v(p) = \begin{cases} \mathbf{D}_L(p), & e_{warp} \leq \tau_{warp} \\ 0, & otherwise \end{cases}, \tag{6}$$

where τ_{warp} is the threshold and is set to 1 in experiments.

Method	Sup	KIT 2012 Out-Noc (%)	KIT 2012 Out-All (%)	KIT 2012 Avg-Noc (px)	KIT 2012 Avg-All (px)	KIT 2015 D1-bg (%)	KIT 2015 D1-fg (%)	KIT 2015 D1-All (%)	Inference Time (s)
DispNet (ICCV'15)	✓	4.11	4.65	0.9	1.0	4.32	4.41	4.34	0.06
GCNet (ICCV'17)	✓	1.77	2.30	0.6	0.7	2.21	6.16	2.87	0.9
SGM-Net (CVPR'17)	✓	2.29	3.50	0.7	0.9	2.66	8.64	3.66	67
AANet (CVPR'20)	✓	1.91	2.42	0.5	0.6	1.99	5.39	2.55	0.06
OASM (ACCV'18)	✗	6.39	8.60	1.3	2.0	6.89	19.42	8.98	0.73
PASMnet_192 (TPAMI'20)	✗	7.14	8.57	1.3	1.5	5.41	16.36	7.23	0.5
F2S (CVPR'20)	✗	4.58	5.11	1.0	1.1	5.01	14.62	6.61	0.05
ChiT‡ (CVPR'22)	✗	-	-	-	-	**2.50**	**5.49**	**3.03**	0.32
DualNet (Step 1)	✗	**2.82**	**3.45**	**0.7**	**0.8**	2.89	8.73	3.86	0.17
DualNet (Step 2)	✗	**2.06**	**2.59**	**0.6**	**0.6**	**2.46**	**5.25**	**2.92**	0.17

Table 1: Comparative results achieved on the KITTI 2012 and KITTI 2015 benchmarks. Sup. indicates whether the method is supervised or not. "-" indicates that results are not available. ‡ denotes that KITTI Eigen Splits (22600 image pairs) and monocular depth priors are used (The best results in **bold**, and the sub-optimal best results in **blue**).

To directly supervise the disparity probability volume, we are motivated from prior works (Zhang et al. 2019), the pseudo-probability distribution $\mathcal{P}$, is generated as follows:

$$\mathcal{P}(d) = \frac{\exp(-c_d^{\mathrm{gt}})}{\sum_{d'=0}^{D-1}\exp(-c_{d'}^{\mathrm{gt}})}, \tag{7}$$

where disparity d is often represented by a floating-point number for sub-pixel matching, $c_d^{gt} = \frac{|d-d_v|}{\delta}$, d_v is the pseudo disparity labels from $\mathbf{D}_v$ and δ>0 is set to 5 in our experiment. D is the maximum disparity.

Student Training. With the pseudo probability distribution $\mathcal{P}$, we can train a student model by forcing its predicted probability distribution $\hat{\mathcal{P}}$ to be similar to $\mathcal{P}$. We use Kullback–Leibler divergence (Hinton, Vinyals, and Dean 2015) to measure the distance between the student model's predicted probability and the pseudo probability. The pseudo-label supervision loss $\mathcal{L}_P$ is defined as:

$$\mathcal{L}_P = \mathcal{L}_{KL}(\hat{\mathcal{P}}||\mathcal{P}) = \sum_{p\in P_v}(\hat{\mathcal{P}}_p)\log(\frac{\hat{\mathcal{P}}_{\mathrm{p}}}{\mathcal{P}_{\mathrm{p}}}). \tag{8}$$

Where P_v represents valid pixels after the left-right check in Eq. 6. In experiments, we find that the trained student model exhibits the capability to become a teacher and further improves the quality of pseudo labels. Therefore, we perform the process of pseudo-label supervision once more.

Experiments

Datasets and Metrics

We train and evaluate our model on four real-world datasets including KITTI 2012 (Geiger, Lenz, and Urtasun 2012), KITTI 2015 (Menze and Geiger 2015), Middlebury (Scharstein et al. 2014), ETH3D (Schöps et al. 2017). Note that our training procedure does not require Ground Truth (GT) for these datasets. Instead, we only use GT for evaluation. For evaluation metrics, end-point error (EPE) and t-pixel error rate (>t px) are adopted. According to evaluated websites, we set t = 1, 2, 3, which defaulting corresponds to the threshold of t = 1 (Bad 1.0) on ETH3D, t = 2 (Bad 2.0) on Middlebury, and t = 3 (Bad 3.0) on KITTI. Besides, we use the density metric to denote the percentage of valid pixels in the generated pseudo labels. Overlap denotes the overlap percentage between the pseudo labels and GT.

Method	Sup.	ETH3D Bad 1.0	ETH3D AvgErr	MID Bad 2.0	MID AvgErr
CFNet (CVPR'21)	✓	3.31	0.24	7.97	2.09
GANet-RSSM (TMM'23)	✓	3.27	0.24	8.21	2.22
UPFNet (TCSVT'23)	✓	3.82	0.25	**5.64**	**1.05**
DualNet (Step 1)	✗	3.91	0.31	14.4	4.79
DualNet (Step 2)	✗	**3.05**	**0.24**	10.8	2.87

Table 2: Quantitative evaluation on ETH3D and Middlebury benchmarks. These supervised methods are fine-tuned on corresponding training sets.

Model	$\mathcal{L}_{photo}$	$\mathcal{L}_s$	$\mathcal{L}_{feat}$	$\mathcal{L}_{aug}$	KIT 2015 EPE	KIT 2015 Bad 3.0	MID EPE	MID Bad 2.0	ETH3D EPE	ETH3D Bad 1.0
Baseline	√	√			0.96	4.10	3.10	12.8	1.39	8.13
DualNet	√	√	√		0.86	3.88	2.64	11.0	0.88	6.32
	√	√	√	√	**0.79**	**3.31**	**1.26**	**9.25**	**0.33**	**3.47**

Table 3: Ablation study on different losses for self-supervised training stage (step 1).

Networks and Training

In this experiment, we evaluate DualNet (step 1& 2) to verify the effectiveness of the whole self-supervised framework. To provide reasonable parameters for finetuning, we adopt the officially provided weights trained on Sceneflow. For step 1, we first train the teacher model with the proposed losses in a self-supervised manner. For step 2, the ar-

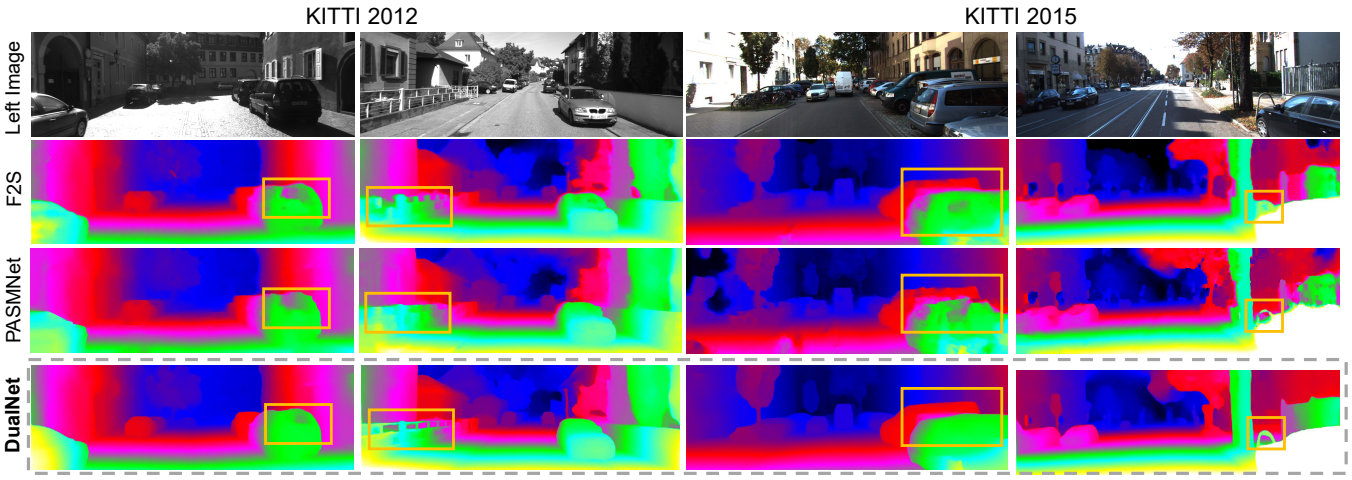

Figure 6: Visualization results achieved by our method and other self-supervised methods from KITTI 2015 & KITTI 2012.

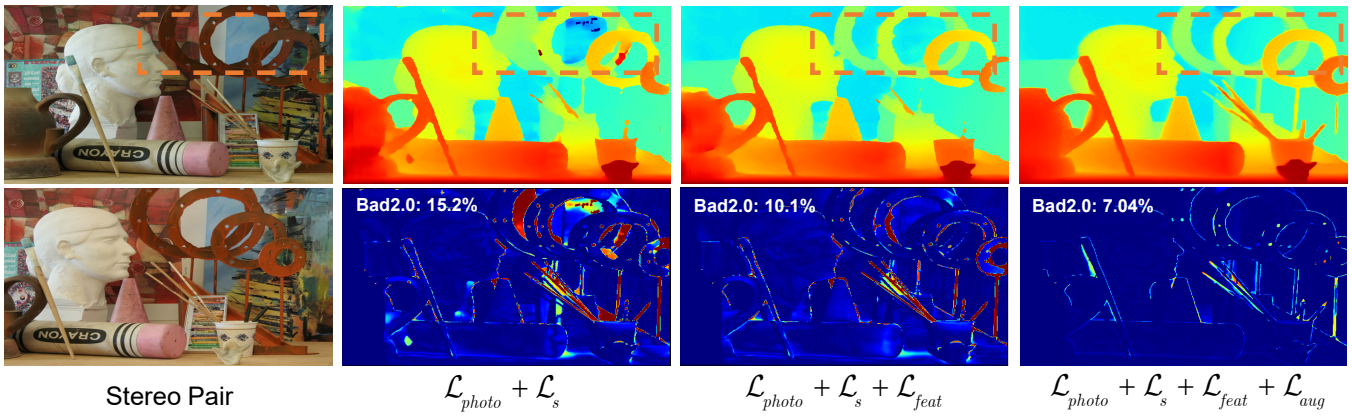

Figure 7: Visual comparison of DualNet on ArtL stereo pair on the Middlebury training dataset. Top row: Overview of generated disparity maps with different loss combinations. Bottom row: the error maps. $\mathcal{L}_{photo}$: Photo-metric Consistency Loss; $\mathcal{L}_{s}$: Disparity Smooth Loss; $\mathcal{L}_{feat}$ Feature-metric Consistency Loss; $\mathcal{L}_{aug}$: Data Augmentation Consistency Loss.

chitecture of the student model is the same as that of the teacher model, while the weights of the student model are re-trained. The generated pseudo labels from the frozen teacher model are modeled as a uni-modal distribution. The output of the student model is then used to align with the teacher's pseudo-probability distribution. Experiments show that performance can be further improved by iterating pseudo-label supervision rounds.

For these two steps, the initial learning rate is set to 1×10^{-4} for the first 20 epochs and decreased to 1×10^{-5} for the remaining 80 epochs. We use 4 NVIDIA A6000 GPU and Pytorch with batch size 8 for all training experiments. The AdamW optimizer with β_1= 0.9 and β_2 = 0.999 is adopted. During training, 320×832 patches are randomly cropped. We use random cropping/color transformation and asymmetric occlusion for data augmentation.

Comparison to State-of-the-art Methods

Unlabeled target data is much easier to obtain than collecting expensive ground-truth disparities. Here, we mainly focus on recent SOTA methods without requiring Ground Truth: self-supervised methods, domain generalization methods, and domain adaption methods. Some supervised ones are also included. We evaluate the final result (step 2) by taking the 2 rounds of pseudo-label supervision strategies.

Self-supervised Methods. Tab. 1 shows that our method outperforms other self-supervised methods by notable margins. The qualitative results are given in Fig. 6, which shows our method excels at recovering challenging regions. Besides, ChiT uses KITTI eigen splits (22600 image pairs) and monocular depth priors for training while our method still outperforms ChiT by only following the regular training setup (Wang et al. 2020; Li and Yuan 2018), using only a mixture of KITTI 12 & 15 image pairs (394 image pairs).

Supervised Methods. Supervised methods can handle challenging regions due to the availability of Ground Truth, exhibiting substantial performance gains in challenging regions. Surprisingly, our self-supervised framework outperforms several supervised SoTA methods, as shown in Tab. 1 and Tab. 2, verifying the superiority of our approach. Note

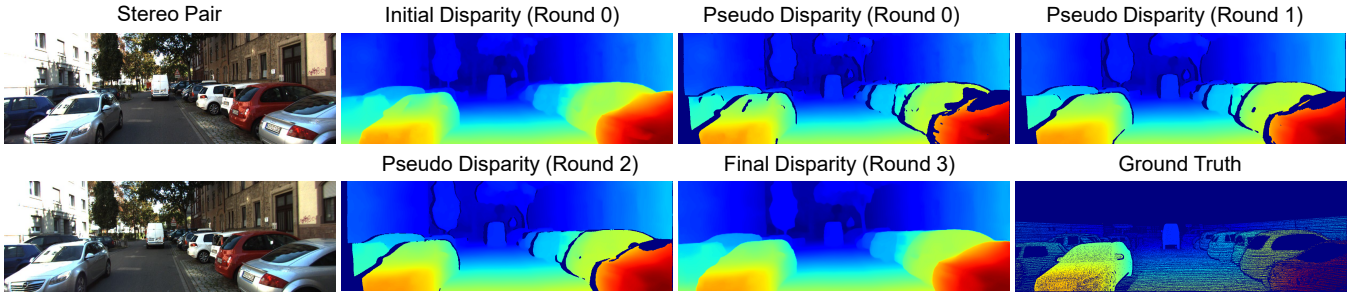

Figure 8: Visualized comparisons between GT disparity and pseudo disparity. Pseudo disparity maps contain high-confidence regions by filtering the outliers from the initial disparity. The threshold τ is set to 1 px.

RND	KIT 2015			KIT 2012		
	Bad 3.0	Dens.	Overlap	Bad 3.0	Dens.	Overlap
0	2.22	82.6	88.3	1.62	74.1	82.6
1	1.70	83.4	89.6	1.44	79.0	85.1
2	**1.63**	**84.2**	**89.9**	**1.22**	**79.1**	**86.1**

Table 4: Quantitative results of pseudo disparity maps at different pseudo-label supervision rounds. **RND 0** denotes that the pseudo disparity maps are generated from the self-supervised teacher learning step. Dens. denotes the density metric.

RND	KIT 2015		MID		ETH3D	
	EPE	Bad 3.0	EPE	Bad 2.0	EPE	Bad 1.0
1	0.70	2.68	1.02	8.00	0.26	2.10
2	**0.66**	2.36	**0.94**	**7.55**	0.22	1.67
3	0.66	**2.32**	0.94	7.64	**0.21**	**1.61**

Table 5: Ablation study on the number of rounds for pseudo-label supervision. The model trained in this round uses pseudo disparity maps generated by the finetuned model from the previous round. GT is not accessible to the model. RND denotes the rounds.

that the Middlebury test benchmark allows only one entry per paper, so we evaluate its training public tables instead.

Ablation Study

Our ablation study is divided into two setups. In the first setup, we perform a self-supervised teacher learning step where the network is trained with the proposed losses. In the second setup, we use pseudo-label supervision, where the network is trained with high-quality pseudo-labels.

Loss Combinations. Tab. 3 shows the effectiveness of proposed loss functions. As can be seen, the performance has gradually improved with the introduction of the proposed losses. A visualization example of the results by our network trained with different losses is shown in Fig. 7. Note that the ArtL stereo image pair has different illumination conditions. If we only adopt the baseline loss $\mathcal{L}_{photo}$ and $\mathcal{L}_s$ for training the model, the predicted disparity map has more blurred areas compared to the results of the proposed loss components. Overall, qualitative and quantitative results demonstrate the effectiveness of the proposed losses.

Number of Pseudo-label Supervision Loops. We find the performance can further be improved by iterative pseudo-label supervision. Here, we show quantitative and qualitative results of the pseudo labels generated by DualNet at each round in Tab. 4 and Fig. 8. Tab. 5 shows that as the number of iterations increases, the overall performance of different datasets improves and gradually tends to be saturated. Therefore, we set the number of rounds to 2 as a trade-off of efficiency and accuracy.

Insights of Effectiveness. We attribute the effectiveness of DualNet to the following four parts. (a) The proposed feature-metric and data augmentation consistency terms lead the model to be more robust, which achieves impressive performance (Bad 3.0: **3.31%**) in the robust self-supervised teacher learning step. (b) Our method can generate accurate pseudo labels with the proposed pseudo-label generation. According to Eq. 6 of the paper, only the high-quality inliers can be kept given a strict threshold. (c) The pseudo labels are semi-dense over GT. The density of GT from KITTI 2012 & 2015 training sets is 28.3% and 19.7%, respectively. Compared with GT, the valid pixels from pseudo labels have denser labels (KITTI 15: $\mathbf{84.2}\%$, KITTI 12: $\mathbf{79.1}\%$) and can provide sufficient geometric information to infer the disparity map. (d) Directly supervising the probability distribution brings performance gain to the student model. Unlike using hard labels such as L1 loss, applying a soft disparity probability distribution directly to cost volumes can reduce its ambiguity (Peng et al. 2022).

Conclusion

In this paper, we present ***DualNet***, a novel self-supervised framework for stereo matching, featuring two main steps: self-supervised teacher learning and pseudo-label supervised student training. For step 1, we introduce a feature-metric consistency term and data augmentation consistency to enhance the model's robustness against common disturbances. For step 2, we incorporate the concept of pseudo-labeling, where the probability distribution of the student model is used to align with the teacher's pseudo-probability distribution. Extensive experiments demonstrate the effectiveness of the proposed self-supervised framework.

Acknowledgements

This work was partly supported by the Shenzhen Science and Technology Program under Grant RCBS20231211090736065, GuangDong Basic and Applied Basic Research Foundation under Grant 2023A151511, Guangdong Natural Science Fund under Grant 2024A1515010252, and Longgang District Shenzhen's "Ten Action Plan" for Supporting Innovation Projects under Grant LGKCSDPT2024002/LGKCSDPT2024003. This work was also supported by the Innohk Initiative, The Government of Hong Kong, SAR (HKSAR), and Laboratory for Artificial Intelligence (AI)-Powered Financial Technologies.

References

Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; and Rodrigo, R. 2022. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9902–9912.

Aleotti, F.; Tosi, F.; Zhang, L.; Poggi, M.; and Mattoccia, S. 2020. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *European Conference on Computer Vision (ECCV)*.

Bangunharcana, A.; Cho, J. W.; Lee, S.; Kweon, I. S.; Kim, K.-S.; and Kim, S. 2021. Correlate-and-Excite: Real-Time Stereo Matching via Guided Cost Volume Excitation. In *2021 IEEE International Conference on Intelligent Robots and Systems (IROS)*, 3542–3548. IEEE.

Ding, Y.; Zhu, Q.; Liu, X.; Yuan, W.; Zhang, H.; and Zhang, C. 2022. Kd-mvs: Knowledge distillation based self-supervised learning for multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 630–646. Springer.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3354–3361.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Li, A.; and Yuan, Z. 2018. Occlusion aware stereo matching via cooperative unsupervised learning. In *Asian Conference on Computer Vision (ACCV)*, 197–213. Springer.

Li, A.; Yuan, Z.; Ling, Y.; Chi, W.; Zhang, S.; and Zhang, C. 2021. Unsupervised occlusion-aware stereo matching with directed disparity smoothing. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 23(7): 7457–7468.

Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; and Liu, S. 2022. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 16263–16272.

Liu, P.; King, I.; Lyu, M. R.; and Xu, J. 2020. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 6648–6657.

Menze, M.; and Geiger, A. 2015. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3061–3070.

Peng, R.; Wang, R.; Wang, Z.; Lai, Y.; and Wang, R. 2022. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8645–8654.

Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; and Westling, P. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition (GCPR)*, 31–42. Springer.

Schöps, T.; Schönberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. 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 (CVPR)*, 2538–2547.

Shen, Z.; and Dai, Y. 2021. CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 13906–13915.

Shen, Z.; Dai, Y.; Song, X.; Rao, Z.; Zhou, D.; and Zhang, L. 2022. PCW-Net: Pyramid combination and warping cost volume for stereo matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 280–297. Springer.

Shen, Z.; Song, X.; Dai, Y.; Zhou, D.; Rao, Z.; and Zhang, L. 2023. Digging Into Uncertainty-Based Pseudo-Label for Robust Stereo Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2).

Su, Q.; and Ji, S. 2022. Chitransformer: Towards reliable stereo from cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1939–1949.

Teed, Z.; and Deng, J. 2020. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 402–419.

Tonioni, A.; Poggi, M.; Mattoccia, S.; and Di Stefano, L. 2019. Unsupervised domain adaptation for depth prediction from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10): 2396–2409.

Wang, L.; Guo, Y.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; and An, W. 2020. Parallax attention for unsupervised stereo correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Wang, X.; Xu, G.; Jia, H.; and Yang, X. 2024a. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 19701–19710.

Wang, Y.; Li, K.; Wang, L.; Hu, J.; Wu, D. O.; and Guo, Y. 2025. ADStereo: Efficient Stereo Matching with Adaptive Downsampling and Disparity Alignment. *IEEE Transactions on Image Processing (TIP)*.

Wang, Y.; Wang, L.; Li, K.; Zhang, Y.; Wu, D. O.; and Guo, Y. 2024b. Cost volume aggregation in stereo matching revisited: A disparity classification perspective. *IEEE Transactions on Image Processing (TIP)*.

Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023. Iterative Geometry Encoding Volume for Stereo Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21919–21928.

Yang, G.; Song, X.; Huang, C.; Deng, Z.; Shi, J.; and Zhou, B. 2019. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 899–908.

Yang, G.; Zhao, H.; Shi, J.; Deng, Z.; and Jia, J. 2018. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European conference on computer vision (ECCV)*, 636–651.

Yen, Y.-T.; Lu, C.-N.; Chiu, W.-C.; and Tsai, Y.-H. 2022. 3D-PL: Domain Adaptive Depth Estimation with 3D-Aware Pseudo-Labeling. In *European Conference on Computer Vision (ECCV)*, 710–728. Springer.

Zhang, F.; Chen, Y.; Li, Z.; Hong, Z.; Liu, J.; Ma, F.; Han, J.; and Ding, E. 2019. Acfnet: Attentional class feature network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 6798–6807.

Zhang, Z.; Zhang, Q.; Li, G.; Luan, J.; Yang, M.; Wang, Y.; and Zhao, L. 2025. DyArtbank: Diverse artistic style transfer via pre-trained stable diffusion and dynamic style prompt Artbank. *Knowledge-Based Systems*, 310: 112959.

Zhang, Z.; Zhang, Q.; Lin, H.; Xing, W.; Mo, J.; Huang, S.; Xie, J.; Li, G.; Luan, J.; Zhao, L.; et al. 2024a. Towards Highly Realistic Artistic Style Transfer via Stable Diffusion with Step-aware and Layer-aware Prompt. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, 7814–7822.

Zhang, Z.; Zhang, Q.; Xing, W.; Li, G.; Zhao, L.; Sun, J.; Lan, Z.; Luan, J.; Huang, Y.; and Lin, H. 2024b. ArtBank: Artistic Style Transfer with Pre-trained Diffusion Model and Implicit Style Prompt Bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7396–7404.

Zhong, Y.; Dai, Y.; and Li, H. 2017. Self-Supervised Learning for Stereo Matching with Self-Improving Ability. *CoRR*, abs/1709.00930.