Geometry-Aware Symmetric Domain Adaptation for Monocular Depth Estimation

Shanshan Zhao[1] Huan Fu[1] Mingming Gong[2,3] Dacheng Tao[1]
[1]UBTECH Sydney AI Center, School of Computer Science, FEIT,
University of Sydney, Darlington, NSW 2008, Australia
[2]Department of Biomedical Informatics, University of Pittsburgh
[3]Department of Philosophy, Carnegie Mellon University
{szha4333@uni., hufu6371@uni., dacheng.tao@}sydney.edu.au mig73@pitt.edu

Abstract

Supervised depth estimation has achieved high accuracy due to the advanced deep network architectures. Since the groundtruth depth labels are hard to obtain, recent methods try to learn depth estimation networks in an unsupervised way by exploring unsupervised cues, which are effective but less reliable than true labels. An emerging way to resolve this dilemma is to transfer knowledge from synthetic images with ground truth depth via domain adaptation techniques. However, these approaches overlook specific geometric structure of the natural images in the target domain (i.e., real data), which is important for high-performing depth prediction. Motivated by the observation, we propose a geometry-aware symmetric domain adaptation framework (GASDA) to explore the labels in the synthetic data and epipolar geometry in the real data jointly. Moreover, by training two image style translators and depth estimators symmetrically in an end-to-end network, our model achieves better image style transfer and generates high-quality depth maps. The experimental results demonstrate the effectiveness of our proposed method and comparable performance against the state-of-the-art. Code will be publicly available at: `https://github.com/sshan-zhao/GASDA`.

1. Introduction

Monocular depth estimation [44, 45, 9, 28] has been an active research area in the field of computer vision. Recent years have witnessed the great strides in this task, especially after deep convolutional neural networks (DCNNs) were exploited to estimate depth from a single image successfully [9]. Until now, there have been lots of follow-up works [35, 30, 8, 31, 54, 51, 10] improving or extending this work. However, since the proposed deep models are trained

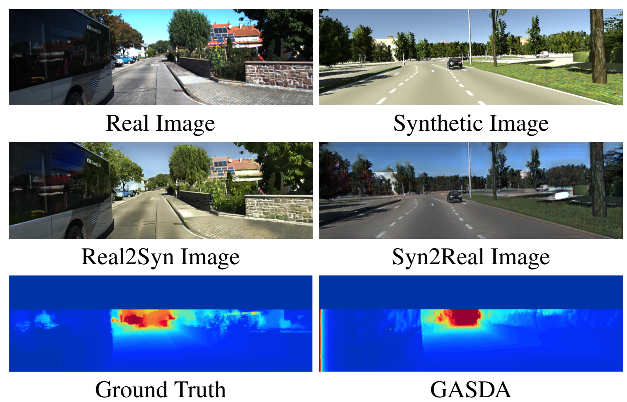


Figure 1: Estimated Depth by GASDA. Top to bottom: input real image in the target domain (KITTI dataset [38]) and synthetic image for training (vKITTI dataset [11]), intermediate generated images in our approach, ground truth depth map and estimated depth map using proposed GASDA.

in a fully supervised fashion, they require a large amount of data with ground truth depth, which is expensive to acquire in practice. To address this issue, unsupervised monocular depth estimation has been proposed [16, 57, 14, 53], using geometry-based cues and without the need of image-depth pairs during training. Unfortunately, this kind of method tends to be vulnerable to illumination change, occlusion and blurring and so on. Compared to real-world data, synthetic data is much easier to obtain the depth map. As a result, some works propose to exploit synthetic data for visual tasks [29, 37, 7]. However, due to domain shift from synthetic to real, the model trained on synthetic data often fails to perform well on real data. To deal with this issue, domain adaptation techniques are utilized to reduce the discrepancy between datasets/domains [1] [2, 5, 37].

[1]We will use *domain* and *dataset* interchangeably for the same meaning in most cases.

Existing works [2, 26, 59] using synthetic data via domain adaptation have achieved impressive performance for monocular depth estimation. These approaches typically perform domain adaptation either based on synthetic-to-realistic translation or inversely. However, due to the lack of paired images, the image translation function usually introduces undesirable distortions in addition to the style change. The distorted image structures significantly degrade the performance of successive depth prediction. Fortunately, the unsupervised cues in the real images, for example, stereo pairs, produces additional constraints on the possible depth predictions. Therefore, it is essential to simultaneously explore both synthetic and real images and the corresponding depth cues for generating higher-quality depth maps.

Motivated by the above analysis, we propose a **Geometry-Aware Symmetric Domain Adaptation Network (GASDA)** for unsupervised monocular depth estimation. This framework consists of two main parts, namely symmetric style translation and monocular depth estimation. Inspired by CycleGAN [61], our GASDA employs both synthetic-to-realistic and realistic-to-synthetic translations coupled with a geometry consistency loss based on the epipolar geomery of the real stereo images. Our network is learned by groundtruth labels from the synthetic domain as well as the epipolar geometry of the real domain. Additionally, the learning process in the real and synthetic domains can be regularized by enforcing consistency on the depth predictions. By training the style translation and depth prediction networks in an end-to-end fashion, our model is able to translate images without distorting the geometric and semantic content, and thus achieves better depth prediction performance. Our contributions can be summarized as follows:

- We propose an end-to-end domain adaptation framework for monocular depth estimation. The model can generate high-quality results for both image style translation and depth estimation.
- We show that training the monocular depth estimator using ground truth depth in the synthetic domain coupled with the epipolar geometry in the real domain can boost the performance.
- We demonstrate the effectiveness of our method on KITTI dataset [38] and the generalization performance on Make3D dataset [45].

2. Related Work

Monocular Depth Estimation has been intensively studied over the past decade due to its crucial role in 3D scene understanding. Typical approaches sought the solution by exploiting probabilistic graphical models (*e.g.*, MRFs) [45, 44, 33], and non-parametric techniques [36, 24, 34]. However, these methods showed some limitations in performance and efficiency because of the employment of hand-crafted features and the low inference speed.

Recent studies demonstrated that high-performing depth estimators can be obtained relying on deep convolutional neural networks (DCNNs) [9, 35, 22, 55, 41, 40, 3, 30, 42, 4]. Eigen *et al.* [9] developed the first end-to-end deep model for depth estimation, which consists of a coarse-scale network and a fine-scale network. To exploit the relationships among image features, Liu *et al.* [35] proposed to integrate continuous CRFs with DCNNs at super-pixel level. While previous works considered depth estimation as a regression task, Fu *et al.* [10] solved depth estimation in the discrete paradigm by proposing an ordinal regression loss to encourage the ordinal competition among depth values.

A weakness of supervised depth estimation is the heavy requirement of annotated training images. To mitigate the issue, several notable attempts have investigated depth estimation in an unsupervised manner by means of stereo correspondence. Xie *et al.* [53] proposed the Deep3D network for 2D-to-3D conversion by minimizing the pixel-wise reconstruction error. This work motivated the development of subsequent unsupervised depth estimation networks [14, 16, 56, 60]. In specific, Garg *et al.* [14] showed that unsupervised depth estimation could be recast as an image reconstruction problem according to the epipolar geometry. Following Garg *et al.* [14], several later works improved the structure by exploiting left-right consistency [16], learning depth in a semi-supervised way [27], and introducing temporal photometric constraints [57].

Domain Adaptation [39] aims to address the problem that the model trained on one dataset fails to generalize to another due to *dataset bias* [49]. In this community, previous works either learn the domain-invariant representations on a feature space [12, 13, 37, 1, 19, 18, 32] or learn a mapping between the source and target domains at feature or pixel level [43, 47, 17, 58]. For example, Long *et al.* [37] aligned feature distribution across the source and target domains by minimizing a Maximum Mean Discrepancy (MMD) [21]. Tzeng *et al.* [50] proposed to minimize MMD and the classification error jointly in a DCNN framework. Sun *et al.* [47] proposed to match the mean and covariance of the two domain's deep features using the Correlation Alignment (CORAL) loss [46].

Coming to domain adaptation for depth estimation, Atapour *et al.* [2] developed a two-stage framework. In specific, they first learned a translator to stylize the natural images so as to make them indistinguishable with the synthetic images, and then trained a depth estimation network using the original synthetic images in a supervised manner. Kundu *et al.* [26] proposed a content congruent regularization method to tackle the model collapse issue caused by domain adaptation in high dimensional feature space. Recently, Zheng

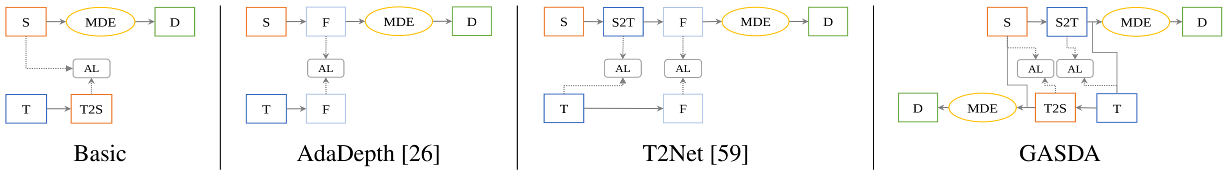


Figure 2: Different frameworks for monocular depth estimation using domain adaptation. Left to right: approach proposed in [26], [59] and this work respectively. S, T, F, S2T (T2S) and D represent the synthetic data, real data, extracted feature, generated data, and estimated depth. AL and MDE mean adversarial loss and monocular depth estimation, respectively. Compared with existing methods, our approach utilizes real stereo data and takes into account synthetic-to-real as well as real-to-synthetic during translation.

et al. [59] developed an end-to-end adaptation network, *i.e.* $\mathrm{T}^2\mathrm{Net}$, where the translation network and the depth estimation network are optimized jointly so that they can improve each other. However, these works overlooked the geometric structure of the natural images from the target domain, which has been demonstrated significant for depth estimation [16, 14]. Motivated by the observation, we propose a novel geometry-aware symmetric domain adaptation network, *i.e.*, GASDA, by exploiting the epipolar geometry of the stereo images. The differences between GASDA and previous depth adaptation approaches [26, 59] are shown in Figure 2.

3. Method

3.1. Method Overview

Given a set of N synthetic image-depth pairs $\{(x_s^i, y_s^i)\}_{i=1}^N$ (*i.e.*, source domain X_s), our goal here is to learn a monocular depth estimation model which can accurately predict depth for natural images contained in X_t (*i.e.*, target domain). It is difficult to guarantee the model generalize well to the real data [2, 59] due to the domain shift. We thus provide a remedy by exploiting the epipolar geometry between stereo images and developing a geometry-aware symmetric domain adaptation network (GASDA). Our GASDA consists of two main parts like existing works, including the style transfer network and the monocular depth estimation network.

Specifically, unlike [2, 59, 26], we consider both synthetic-to-real [59] and real-to-synthetic translations [2, 26]. As a result, we can train two depth estimators F_s and F_t on the original synthetic data (X_s) and the generated realistic data ($G_{s2t}(X_s)$) using the generator G_{s2t} in supervised manners, respectively. These two models are complementary, since F_s has clean training set X_s but dirty test set $G_{t2s}(X_t)$ generated by the generator G_{t2s} with noises, such as distortion and blurs, caused by unsatisfied translation, and vise verse for F_t. Nevertheless, because the depth information is rather relevant to specific scene geometry which might be different between source and target domains, the models trained on X_s or $G_{s2t}(X_s)$ still could fail to perform well on $G_{t2s}(X_t)$ or X_t. To provide a solution, we exploit the epipolar geometry of real stereo pairs $\{(x_{t_l}^i, x_{t_r}^i)\}_{i=1}^M$ ($x_{t_l}^i$ and $x_{t_r}^i$ represent the left and right image respectively[2]) during training to encourage F_t and F_s to capture the relevant geometric structure of target/real data. In addition, we introduce an additional depth consistency loss to enforce the predictions from F_t and F_s are consistent in local regions. The overall framework of GASDA is illustrated in Figure 3. For simplicity, we will omit the superscript i in most cases.

3.2. GASDA

Bidirectional Style Transfer Loss Our goal here is to learn the bidirectional translators G_{s2t} and G_{t2s} to bridge the gap between the source domain (synthetic) X_s and the target domain (real) X_t. Specifically, taking G_{s2t} as an example, we expect the $G_{s2t}(x_s)$ to be indistinguishable from real images in X_t. We thus employ a discriminator D_t, and train G_{s2t} and D_t in an adversarial fashion by performing a minimax game following [20]. The adversarial losses are expressed as:

$$\begin{aligned}\mathcal{L}_{gan}(G_{s2t}, D_t, X_t, X_s) =& \mathbb{E}_{x_t \sim X_t}[D_t(x_t) - 1] + \\ & \mathbb{E}_{x_s \sim X_s}[D_t(G_{s2t}(x_s))], \\ \mathcal{L}_{gan}(G_{t2s}, D_s, X_t, X_s) =& \mathbb{E}_{x_s \sim X_s}[D_s(x_s) - 1] + \\ & \mathbb{E}_{x_t \sim X_t}[D_s(G_{t2s}(x_t))].\end{aligned} \tag{1}$$

Unluckily, the vanilla GANs suffer from mode collapse. To provide a remedy and ensure the input images and the output images paired up in a meaningful way, we utilize the cycle-consistency loss [61]. Specifically, when feeding an image x_s to G_{s2t} and G_{t2s} orderly, the output should be a reconstruction of x_s, and vice versa for x_t, *i.e.* $G_{t2s}(G_{s2t}(x_s)) \approx x_s$ and $G_{s2t}(G_{t2s}(x_t)) \approx x_t$. The cycle consistency loss has the form as:

$$\begin{aligned}\mathcal{L}_{cyc}(G_{t2s}, G_{s2t}) = & \mathbb{E}_{x_s \sim X_s}[||G_{t2s}(G_{s2t}(x_s)) - x_s||_1] \\ & + \mathbb{E}_{x_t \sim X_t}[||G_{s2t}(G_{t2s}(x_t)) - x_t||_1].\end{aligned} \tag{2}$$

Apart from the adversarial loss and cycle consistency loss, we also employ an identity mapping loss [48] to encourage the generators to preserve geometric content. The

[2] We will omit the subscript l of t_l for the left image in most cases.

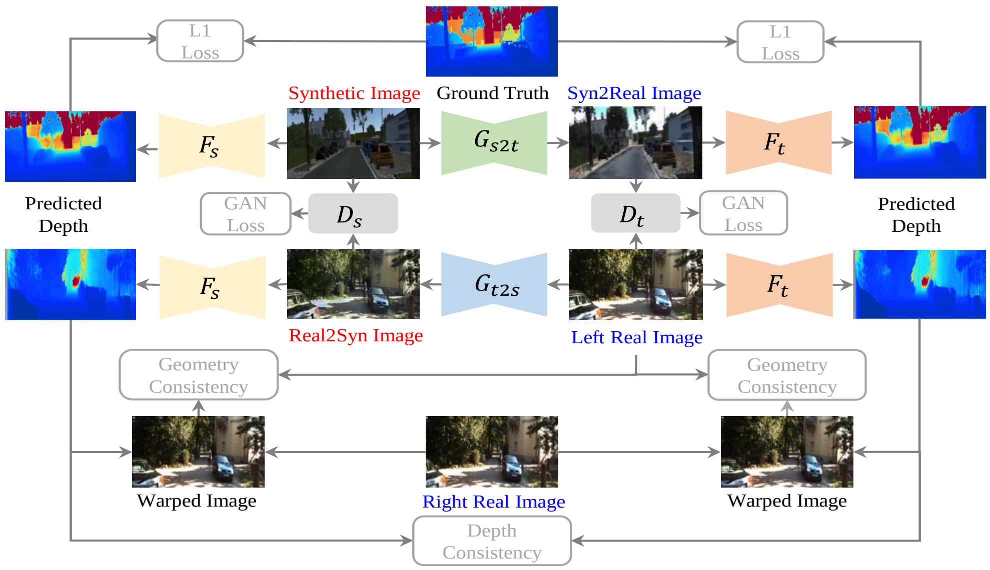


Figure 3: The proposed framework in this paper. It consists of two main parts: image style translation and monocular depth estimation. i) Style translation network, incorporating two generators (*i.e.*, G_{s2t} and G_{t2s}) and two discriminators (*i.e.*, D_t and D_s), is based on CycleGAN [61]. ii) Monocular depth estimation network contains two complementary sub-networks (*i.e.*, F_s and F_t). We omit the side outputs, for brevity. More details can be found in Section 3, Section 4.1.

identity mapping loss is given by:

$$\begin{aligned}\mathcal{L}_{idt}(G_{t2s}, G_{s2t}, X_s, X_t) = \mathbb{E}_{x_s \sim X_s}[||G_{t2s}(x_s) - x_s||_1] \\ + \mathbb{E}_{x_t \sim X_t}[||G_{s2t}(x_t) - x_t||_1].\end{aligned} \tag{3}$$

The full objective for the bidirectional style transfer is as follow:

$$\begin{aligned}\mathcal{L}_{trans}(G_{t2s}, G_{s2t}, D_t, D_s) = {} & \mathcal{L}_{gan}(G_{s2t}, D_t, X_t, X_s) \\ & + \mathcal{L}_{gan}(G_{t2s}, D_s, X_t, X_s) \\ & + \lambda_1 \mathcal{L}_{cyc}(G_{t2s}, G_{s2t}) \\ & + \lambda_2 \mathcal{L}_{idt}(G_{t2s}, G_{s2t}, X_t, X_s)\end{aligned} \tag{4}$$

where λ_1 and λ_2 are the trade-off parameters.

Depth Estimation Loss We can now render the synthetic images to the "style" of the target domain (KITTI), and then capture a new dataset $X_{s2t} = G_{s2t}(X_s)$. We train a depth estimation network F_t on X_{s2t} in a supervised manner using the provided ground truth depth in the synthetic domain X_s. Here, we minimize the ℓ_1 distance between the predicted depth $\tilde{y}_{ts}$ and ground truth depth y_s:

$$\mathcal{L}_{tde}(F_t, G_{s2t}) = ||y_s - \tilde{y}_{ts}||. \tag{5}$$

In addition to F_t, we also train a complementary depth estimator F_s on X_s directly with the ℓ_1 loss:

$$\mathcal{L}_{sde}(F_s) = ||y_s - \tilde{y}_{ss}|| \tag{6}$$

where $\tilde{y}_{ss} = F_s(x_s)$ is the output of F_s. Both the F_s and F_t are important backbones to alleviate the issue of geometry and semantic inconsistency coupled with the subsequent losses. The full depth estimation loss is expressed as:

$$\mathcal{L}_{de}(F_t, F_s, G_{s2t}) = \mathcal{L}_{sde}(F_s) + \mathcal{L}_{tde}(F_t, G_{s2t}). \tag{7}$$

Geometry Consistency Loss Combining the components above, we have already formulated a naive depth adversarial adaptation framework. However, the G_{s2t} and G_{t2s} are usually imperfect, which would make the predictions $\tilde{y}_{st} = F_s(G_{t2s}(x_t))$ and $\tilde{y}_{tt} = F_t(x_t)$ unsatisfied. Besides, previous depth adaptation approaches overlook the specific physical geometric structure which may vary from scenes/datasets. Our main objective is to accurately estimate depth for real scenes, so we consider the geometric structure of the target data in the training phase. To this end, we present a geometric constraint on F_t and F_s by exploiting the epipolar geometry of real stereo images and unsupervised cues. Specifically, we generate an inverse warped image from the right image using the predicted depth, to reconstruct the left. We thus combine an ℓ_1 with single scale SSIM [52] term as the geometry consistency loss to align the stereo images:

$$\begin{aligned}&\mathcal{L}_{tgc}(F_t) = \eta \frac{1 - SSIM(x_t, x'_{tt})}{2} + \mu||x_t - x'_{tt}||, \\ &\mathcal{L}_{sgc}(F_s, G_{t2s}) = \eta \frac{1 - SSIM(x_t, x'_{st})}{2} + \mu||x_t - x'_{st}||, \\ &\mathcal{L}_{gc}(F_t, F_s, G_{t2s}) = \mathcal{L}_{tgc}(F_t) + \mathcal{L}_{sgc}(F_s, G_{t2s})\end{aligned} \tag{8}$$

where $\mathcal{L}_{gc}$ represents the full geometry consistency loss, $\mathcal{L}_{tgc}$ and $\mathcal{L}_{sgc}$ denote the geometry consistency loss of F_t

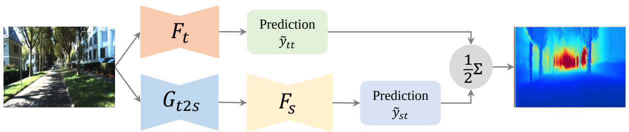


Figure 4: Inference Phase (Section 3.3).

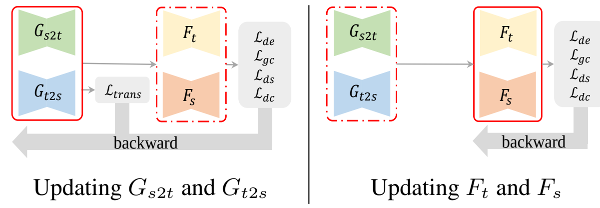


Figure 6: Iteratively updating stage. We learn our model by iteratively updating image style translators and depth estimators, *i.e.*, freezing the module with dashed box while updating the one with solidline box. See main text for details. We omit D_t and D_s for brevity.

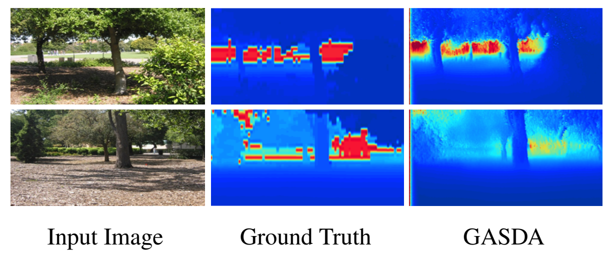


Figure 8: Qualitative results on Make3D dataset [45]. Left to right: input image, ground truth depth, and our result.

and F_s respectively. $x_{tt}^{'}$ ($x_{st}^{'}$) is the inverse warp of x_{t_r} using bilinear sampling [23] based on the estimated depth map y_{tt} (y_{st}), the baseline distance between the cameras and the camera focal length [16]. In our experiments, η is set to be 0.85, and μ is 0.15.

Depth Smoothness Loss To encourage depths to be consistent in local homogeneous regions, we exploit an edge-aware depth smoothness loss:

$$\mathcal{L}_{ds}(F_t, F_s, G_{t2s}) = e^{-\nabla x_t}||\nabla \tilde{y}_{tt}|| + e^{-\nabla x_t}||\nabla \tilde{y}_{st}|| \quad (9)$$

where ∇ is the first derivative along spatial directions. We only apply the smoothness loss to X_t and X_{t2s} (real data), since X_s and X_{s2t} (synthetic data) have full supervision.

Depth Consistency Loss We find that the predictions for x_t, *i.e.*, $F_t(x_t)$ and $F_s(G_{t2s}(x_t))$, show inconsistency in many regions, which is in contrast to our intuition. One of the possible reason is that G_{t2s} might fail to translate x_t with details. To enforce such coherence, we introduce an ℓ_1 depth consistency loss with respect to $\tilde{y}_{tt}$ and $\tilde{y}_{st}$ as follows:

$$\mathcal{L}_{dc}(F_t, F_s, G_{t2s}) = ||\tilde{y}_{tt} - \tilde{y}_{st}||. \quad (10)$$

Full Objective Our final loss function has the form as:

$$\begin{aligned}
&\mathcal{L}(G_{s2t}, G_{t2s}, D_t, D_s, F_t, F_s) \\
&= \mathcal{L}_{trans}(G_{s2t}, G_{t2s}, D_t, D_s) + \gamma_1 \mathcal{L}_{de}(F_t, F_s, G_{s2t}) \\
&+ \gamma_2 \mathcal{L}_{gc}(F_t, F_s, G_{t2s}) + \gamma_3 \mathcal{L}_{dc}(F_t, F_s, G_{t2s}) \\
&+ \gamma_4 \mathcal{L}_{ds}(F_t, F_s, G_{t2s})
\end{aligned} \quad (11)$$

where $\gamma_n (n \in \{1, 2, 3, 4\})$ are trade-off factors. We optimize this objective function in an end-to-end deep network.

3.3. Inference

In the inference phase, we aim to predict the depth map for a given image in real domain (*e.g.* KITTI dataset [38]) using the resultant models. In fact, there are two paths acquiring predicted depth maps: $x_t \rightarrow F_t(x_t) \rightarrow \tilde{y}_{tt}$ and $x_t \rightarrow G_{t2s}(x_t) \rightarrow x_{t2s} \rightarrow F_s(x_{t2s}) \rightarrow \tilde{y}_{st}$, as shown in Figure 4, and the final prediction is the average of $\tilde{y}_{tt}$ and $\tilde{y}_{st}$:

$$\tilde{y}_t = \frac{1}{2}(\tilde{y}_{tt} + \tilde{y}_{st}). \quad (12)$$

4. Experiments

In this section, we first present the details about our network architecture and the learning strategy. Then, we perform GASDA on one of the largest dataset in the context of autonomous driving, *i.e.*, KITTI dataset [38]. We also demonstrate the generalization capabilities of our model to other real-world scenes contained in Make3D [45]. Finally, we conduct various ablations to analyze GASDA.

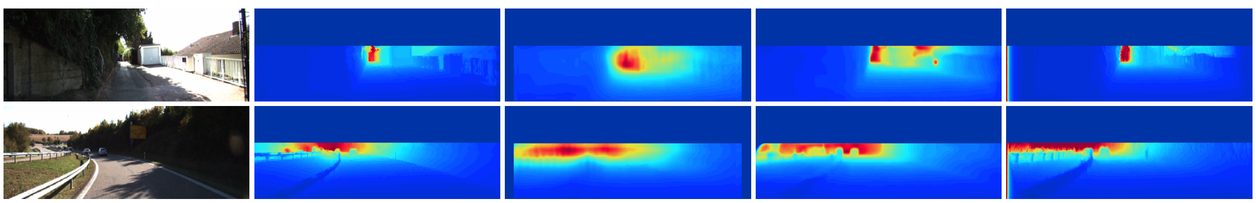
4.1. Implementation Details

Network Architecture Our proposed framework consists of six sub-networks, which can be divided into three groups: G_{s2t} and G_{t2s} for image style translation, D_t and D_s for discrimination, F_t and F_s for monocular depth estimation. The networks in each group share the identical network architecture but are with different parameters. Specifically, we employ generators (G_{s2t} and G_{t2s}) and discriminators (D_s and D_t) provided by CycleGAN [61]. For monocular depth estimators F_t and F_s, we utilize the standard encoder-decoder structures with skip-connections and side outputs as [59].

Datasets The target domain is KITTI [38], which is a real-world computer vision benchmark consisting of $42,382$ rectified stereo pairs in the resolution about 375×1242. In our experiments, the ground truth depth maps provided by KITTI are only for evaluation purpose. The source domain is Virtual KITTI (vKITTI) [11], which contains 50 photo-realistic synthetic videos with $21,260$ image-depth pairs of

Method	Supervised	Dataset	Cap	Error Metrics (lower, better)				Accuracy Metrics (higher, better)		
				Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [9]	Yes	K	$80m$	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.* [35]	Yes	K	$80m$	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Zhou *et al.* [60]	No	K	$80m$	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhou *et al.* [60]	No	K+CS	$80m$	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Kuznietsov *et al.* [27]	Semi	K	$80m$	0.113	0.741	4.621	0.189	0.862	0.960	0.986
Godard *et al.* [16]	No	K	$80m$	0.148	1.344	5.927	0.247	0.803	0.922	0.964
All synthetic(baseline1)	No	S	$80m$	0.253	2.303	6.953	0.328	0.635	0.856	0.937
All real(baseline2)	No	K	$80m$	0.158	1.151	5.285	0.238	0.811	0.934	0.970
Kundu *et al.* [26]	No	K+S(DA)	$80m$	0.214	1.932	7.157	0.295	0.665	0.882	0.950
Kundu *et al.* [26]	Semi	K+S(DA)	$80m$	0.167	1.257	5.578	0.237	0.771	0.922	0.971
GASDA	No	K+S(DA)	$80m$	**0.149**	**1.003**	**4.995**	**0.227**	**0.824**	**0.941**	**0.973**
Kuznietsov *et al.* [27]	Yes	K	$50m$	0.117	0.597	3.531	0.183	0.861	0.964	0.989
Garg *et al.* [14]	No	K	$50m$	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Godard *et al.* [16]	No	K	$50m$	0.140	0.976	4.471	0.232	0.818	0.931	0.969
All synthetic(baseline1)	No	S	$50m$	0.244	1.771	5.354	0.313	0.647	0.866	0.943
All real(baseline2)	No	K	$50m$	0.151	0.856	4.043	0.227	0.824	0.940	0.973
Kundu *et al.* [26]	No	K+S(DA)	$50m$	0.203	1.734	6.251	0.284	0.687	0.899	0.958
Kundu *et al.* [26]	Semi	K+S(DA)	$50m$	0.162	1.041	4.344	0.225	0.784	0.930	0.974
Zheng *et al.* [59]	No	K+S(DA)	$50m$	0.168	1.199	4.674	0.243	0.772	0.912	0.966
GASDA	No	K+S(DA)	$50m$	**0.143**	**0.756**	**3.846**	**0.217**	**0.836**	**0.946**	**0.976**

Table 1: Results on KITTI dataset using the test split suggested in [9]. For the training data, K represents KITTI dataset, CS is CityScapes dataset [6], and S is vKITTI dataset. Methods, which apply domain adaptation techniques, are marked by the gray.



Input Image Ground Truth Eigen *et.al.* [9] Zheng *et.al.* [59] GASDA

Figure 5: Qualitative comparison of our results against methods proposed by Eigen *et al.* [9] and Zheng *et al.* [59] on KITTI. Ground truth has been interpolated for visualization. To facilitate comparison, we mask out the top regions, where ground truth depth is not available. Our approach preserves more details and yields high-quality depth maps.

size 375×1242. Additionally, in order to study the generalization performance of our approach, we also apply the trained model to Make3D dataset [45]. Since Make3D does not offer stereo images, we directly evaluate our model on the test split without training or further fine-tuning.

Training Details We implement GASDA in *PyTorch*. We train our model in a two-stage manner, *i.e.*, a warming up stage and end-to-end iteratively updating stage. In the warming up stage, we first optimize the style transfer networks for 10 epochs with the momentum of $\beta_1 = 0.5$, $\beta_2 = 0.999$, and the initial learning rate of $\alpha = 0.0002$ using the ADAM solver [25]. Then we train F_t on $\{X_t, G_{s2t}(X_s)\}$, and F_s on $\{X_s, G_{t2s}(X_t)\}$ for around 20 epochs by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\alpha = 0.0001$. To make style translators generate high-quality images, so as to improve the subsequent depth estimators, we fine-tune the network in an end-to-end iteratively updating fashion as shown in Figure 6. In specific, we optimize G_{s2t} and G_{t2s} with the supervision of F_t and F_s for m epochs, and then train F_s and F_t for n epochs. We set $m = 3$ and $n = 7$ in our experiments, and repeat this process until the network converges (around 40 epochs). In this stage, we employ the same momentum and solver as the first stage with the learning rates of $2e-6$ and $1e-5$ for the two respectively. The trade-off factors are set to $\lambda_1 = 10$, $\lambda_2 = 30$, $\gamma_1 = 50$, $\gamma_2 = 50$ and $\gamma_3 = 50$ and $\gamma_4 = 0.5$. In the training phase, we downsample all the images to 192×640, and increase the training set size using some common data augmentation strategies, including random horizontal flipping, rotation with the degrees of $[-5°, 5°]$, and brightness adjustment.

4.2. KITTI Dataset

We test our models on the 697 images extracted from 29 scenes, and use all the $23,488$ images contained in other 32 scenes for training $(22,600)$ and validation (888) [9, 16]. To make a comparison with previous works, we evaluate our results in the regions with the ground truth depth less than $80m$ or $50m$ using standard error and accuracy metrics [16, 59]. Note that, the maximum depth value in vKITTI is $655.35m$ instead of $80m$ in KITTI, but unlike [59], we do not clip the depth maps of vKITTI to $80m$ during training. In Table 1, we report the benchmark scores on the Eigen s-

Method	Supervised	Dataset	Error Metrics (lower, better)				Accuracy Metrics (higher, better)		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard *et al.* [16]	No	K	0.124	1.388	6.125	0.217	0.841	0.936	0.975
Godard *et al.* [16]	No	K+CS	0.104	1.070	5.417	0.188	0.875	0.956	0.983
Atapour *et al.* [2]	No	K+S*(DA)	**0.101**	1.048	5.308	0.184	**0.903**	**0.988**	**0.992**
GASDA	No	K+S(DA)	0.106	**0.987**	**5.215**	**0.176**	0.885	0.963	0.986

Table 2: Results on 200 training images of KITTI stereo 2015 benchmark [15]. S^* is captured from GTA5, and more similar to real data than vKITTI. Our approach yields lower errors than state-of-the-art approaches, and achieve competitive accuracy compared with [2].

Real Image CycleGAN [61] GASDA | Synthetic Image CycleGAN [61] GASDA

Figure 7: Qualitative image style translation results of our approach and CycleGAN [61]. Left: real-to-synthetic translation; Right: synthetic-to-real translation. Our method can preserve geometric and semantic content better for both synthetic-to-real translation and the inverse one. Note that, the translation result is a by-product of GASDA. The improvement is marked by the yellow box.

Method	Trained*	Error Metrics (lower, better)		
		Abs Rel	Sq Rel	RMSE
Karsch *et al.* [24]	Yes	0.398	4.723	7.801
Laina *et al.* [30]	Yes	0.198	1.665	5.461
Kundu *et al.* [26]	Yes	0.452	5.71	9.559
Godard *et al.* [16]	No	0.505	10.172	10.936
Kundu *et al.* [26]	No	0.647	12.341	11.567
Atapour *et al.* [2]	No	0.423	9.343	**9.002**
GASDA	No	**0.403**	**6.709**	10.424

Table 4: Results on 134 test images of Make3D [45]. Trained* indicates whether the model is trained on Make3D or not. Errors are computed for depths less than $70m$ in a central image crop [16]. It can be observed that our approach is comparable with those trained on Make3D.

plit [9] where the training sets are only KITTI and vKITTI. GASDA obtains a convincible improvement over previous state-of-the-art methods. Specifically, we make the comparisons with two baselines, *i.e.,* All synthetic (baseline1, trained on labeled synthetic data) and All real (baseline2, trained on real stereo pairs), and the latest domain adaptation methods [59, 26] and (semi-)supervised/unsupervised methods [9, 35, 27, 14, 16, 60]. The significant improvements in all the metrics demonstrate the superiority of our method. Note that, GASDA yields higher scores than [26] which employs additional ground truth depth maps for natural images contained in KITTI. GASDA cannot outperform [2] in the Eigen split. The main reason is that the synthetic images employed in [2] are captured from GTA5 [3], and the domain shift between GTA5 and KITTI is not that significant than the one between vKITTI and KITTI.

[3]https://github.com/aitorzip/DeepGTAV.

In addition, the training set size in [2] is about three times than ours. However, GASDA performs competitively on the official KITTI stereo 2015 dataset and Make3D compared with [2], as reported in Table 2 and Table 4. Apart from quantitative results, we also show some example outputs in Figure 5. Our approach preserves more details, and is able to recover depth information of small objects, such as the distant cars and rails, and generate clear boundaries.

4.3. Make3D Dataset

To discuss the generalization capabilities of GASDA, we evaluate our approach on Make3D dataset [45] quantitatively and qualitatively. We do not train or further fine-tune our model using the images provide by Make3D. As shown in Table 4 and Figure 8, although the domain shift between Make3D and KITTI is large, our model still performs well. Compared with state-of-the-art models [26, 24, 30] trained on Make3D in a supervised manner and others using domain adaptation [26, 2], GASDA obtains impressive performance.

4.4. Ablation Study

Here, we conduct a series of ablations to analyze our approach. Quantitative results are shown in Table 3, and some sampled results for style transfer are shown in Figure 7.

Domain Adaptation We first demonstrate the effectiveness of domain adaptation by comparing two simple models, *i.e.* SYN (F_s trained on X_s) and SYN2REAL (F_t trained on $G_{s2t}(X_s)$). As shown in Table 3, SYN cannot capture satisfied scores on KITTI due to the domain shift. After the translation, the domain shift is reduced which means that the synthetic data distribution is relative closer to real data

Method	Error Metrics (lower, better)				Accuracy Metrics (higher, better)		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Domain Adaptation							
SYN	0.253	2.303	6.953	0.328	0.635	0.856	0.937
SYN2REAL	0.229	2.094	6.530	0.294	0.691	0.886	0.951
SYN2REAL-E2E	**0.220**	**1.969**	**6.377**	**0.284**	**0.703**	**0.895**	**0.956**
Geometry Consistency							
REAL	0.158	1.151	5.285	0.238	0.811	0.934	0.970
SYN-GC	0.156	1.123	5.255	0.235	0.814	0.937	0.971
SYN2REAL-GC	0.153	**1.112**	**5.213**	0.233	0.819	0.938	0.972
SYN2REAL-GC-E2E	**0.152**	1.130	5.227	**0.231**	**0.821**	**0.939**	**0.972**
Symmetric Domain Adaptation							
REAL2SYN-SYN-GC-E2E	0.160	1.226	5.412	0.240	0.806	0.933	0.969
GASDA-w/oDC	0.151	1.098	5.136	0.230	0.822	0.940	0.972
GASDA-F_t	0.150	1.014	5.041	0.228	**0.824**	**0.941**	**0.973**
GASDA-F_s	0.156	1.087	5.157	0.235	0.813	0.936	0.971
GASDA	**0.149**	**1.003**	**4.995**	**0.227**	**0.824**	**0.941**	**0.973**

Table 3: Quantitative results for ablation study on KITTI dataset using the test split suggested in [9]. SYN, REAL, REAL2SYN, and SYN2REAL represent the model trained on X_s, X_t, $G_{t2s}(X_t)$, and $G_{s2t}(X_s)$; E2E represents the end-to-end training; GC and DC denote the geometry consistency and depth consistency, respectively; GASDA-F_t (F_s) represents the output of F_t (F_s) in GASDA.

distribution. Thus, SYN2REAL is able to generalize better to real images. Further, we train the style translators (G_{s2t} and G_{t2s}) and the depth estimation network (F_t) in an end-to-end fashion (SYN2REAL-E2E), which guides to a further improvement as compared to SYN2REAL. As a conclusion, the depth estimation network can improve the style transfer by providing a pixel-wise semantic constraint to the translation networks. Moreover, we can also observe the improvement in Figure 7 by comparing the translation results of original CycleGAN [61] with ours.

Geometry Consistency We then study the significance of the geometric constraint coming from stereo images based on the epipolar geometry. In specific, we employ the stereo images provided by KITTI when optimizing F_t in SYN2REAL-E2E. We enforce the geometry consistency between the stereo images as a constraint as stated in Eq. 8. The model SYN2REAL-GC-E2E outperforms SYN2REAL-E2E by a large margin, which demonstrates that the geometry consistency constraint can significantly improve standard domain adaptation frameworks. On the other hand, the comparisons among SYN2REAL-GC, SYN-GC (trained on real data and synthetic data without domain adaptation) and REAL (F_t trained on real stereo images without extra data) can show the significance of synthetic data with ground truth depth and domain adaptation.

Symmetric Domain Adaptation In contrast to previous works, we expect to fully take advantage of the bidirectional style translators G_{s2t} and G_{t2s}. Thus, we learn REAL2SYN-SYN-GC-E2E whose network architecture is symmetrical to the aforementioned SYN2REAL-GC-E2E. We jointly optimized the two coupled with a depth consistency loss. As shown in Table 3, GASDA is superior than GASDA-w/oDC which demonstrates the effectiveness of the depth consistency loss. In addition, the comparisons (GASDA-F_t *v.s.* SYN2ERAL-GC-E2E and GASDA-F_s *v.s.* REAL2SYN-GC-E2E) show that the two can benefit each other in the jointly training.

5. Conclusion

In this paper, we present an unsupervised monocular depth estimation framework GASDA, which trains the monocular depth estimation model using the labelled synthetic data coupled with the epipolar geometry of real stereo data in a unified and symmetric deep learning network. Our main motivation is learning a depth estimation model from synthetic image-depth pairs in a supervised fashion, and at the same time taking into account the specific scene geometry information of the target data. Moreover, to alleviate the issues caused by domain shift, we reduce the domain discrepancy using the bidirectional image style transfer. Finally, we implement image translation and depth estimation in an end-to-end network so that then can improve each other. Experiments on KITTI and Make3D datasets show GASDA is able to generate desirable results quantitatively and qualitatively, and generalize well to unseen datasets.

6. Acknowledgement

This research was supported by Australian Research Council Projects FL-170100117, DP-180103424 and IH-180100002.

References

[1] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

[2] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 18, page 1, 2018.

[3] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *arXiv preprint arXiv:1605.02305*, 2016.

[4] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738, 2016.

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[7] 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.

[8] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[10] 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.

[11] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[12] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[14] 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.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012.

[16] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017.

[17] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012.

[18] Mingming Gong, Kun Zhang, Biwei Huang, Clark Glymour, Dacheng Tao, and Kayhan Batmanghelich. Causal generative domain adaptation networks. *arXiv preprint arXiv:1804.04333*, 2018.

[19] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848, 2016.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[21] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[22] Lei He, Guanghui Wang, and Zhanyi Hu. Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing*, 2018.

[23] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[24] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2144–2158, 2014.

[25] Diederik P Kingma and Jimmy Lei Ba. Adam: Amethod for stochastic optimization. In *Proc. 3rd Int. Conf. Learn. Representations*, 2014.

[26] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. *arXiv preprint arXiv:1803.01599*, 2018.

[27] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6647–6655, 2017.

[28] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014.

[29] Wei-Sheng Lai, Jia-Bin Huang, and Ming-Hsuan Yang. Semi-supervised learning for optical flow with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 354–364, 2017.

[30] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[31] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.

[32] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[33] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–1260. IEEE, 2010.

[34] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2011.

[35] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian D Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):2024–2039, 2016.

[36] Miaomiao Liu, Mathieu Salzmann, and Xuming He. Discrete-continuous depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723, 2014.

[37] Mingsheng Long, Guiguang Ding, Jianmin Wang, Jiaguang Sun, Yuchen Guo, and Philip S Yu. Transfer sparse coding for robust image representation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 407–414, 2013.

[38] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015.

[39] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[40] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018.

[41] Vamshi Krishna Repala and Shiv Ram Dubey. Dual cnn models for unsupervised monocular depth estimation. *arXiv preprint arXiv:1804.06324*, 2018.

[42] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5506–5514, 2016.

[43] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[44] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.

[45] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009.

[46] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, volume 6, page 8, 2016.

[47] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[48] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[49] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.

[50] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[51] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2015.

[52] 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.

[53] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857. Springer, 2016.

[54] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of CVPR*, volume 1, 2017.

[55] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3917–3925, 2018.

[56] 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 (CVPR)*, volume 2, 2018.

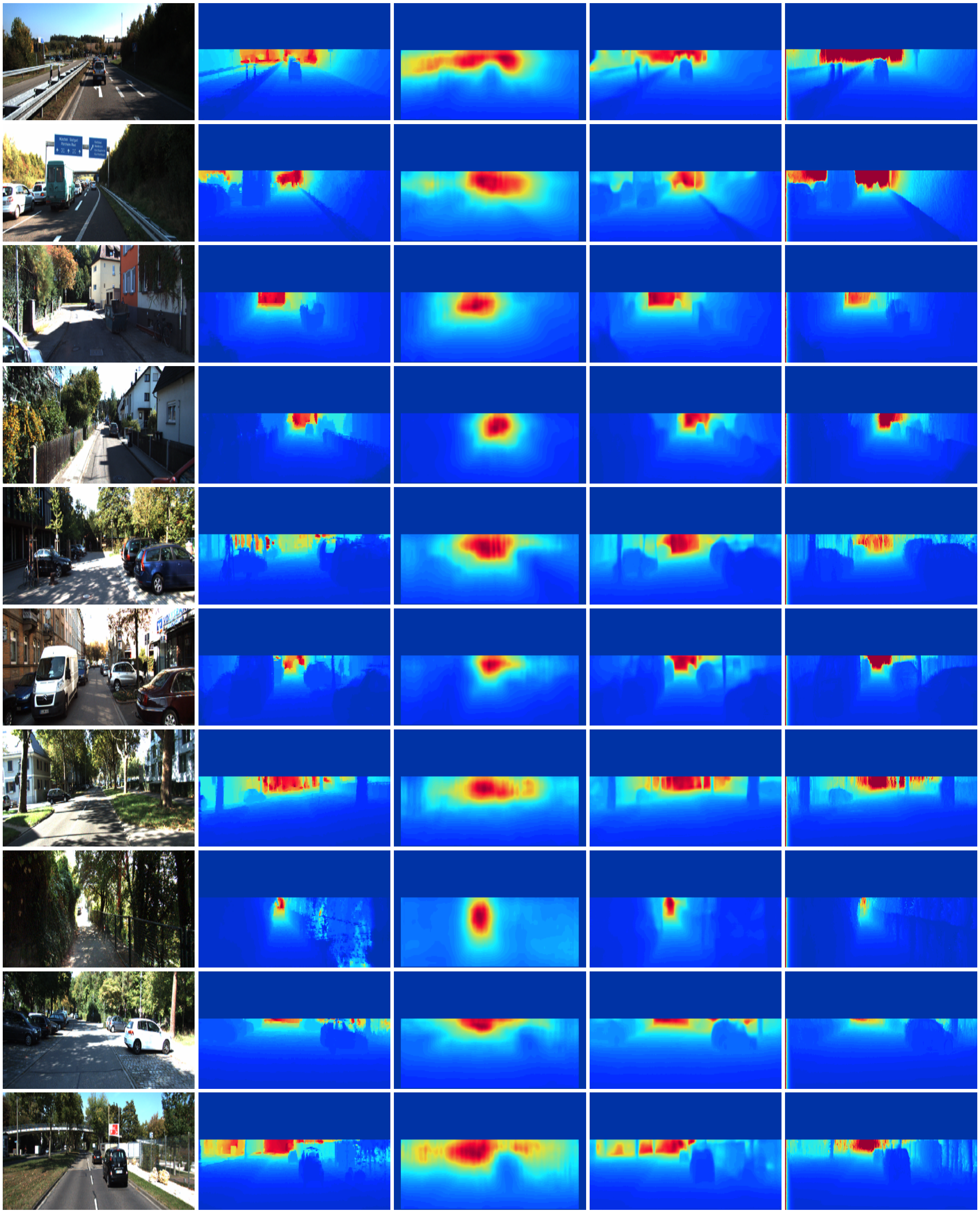
[57] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018.

[58] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

[59] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[60] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017.

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.



Input Image Ground Truth Eigen *et.al.* [9] Zheng *et.al.* [59] GASDA

Figure 9: Qualitative comparisons of our results with methods proposed by Eigen *et al.* [9] and Zheng *et al.* [59] on theKITTI Eigen Split [9]. The model is trained on KITTI using the split of Eigen *et al.* [9].

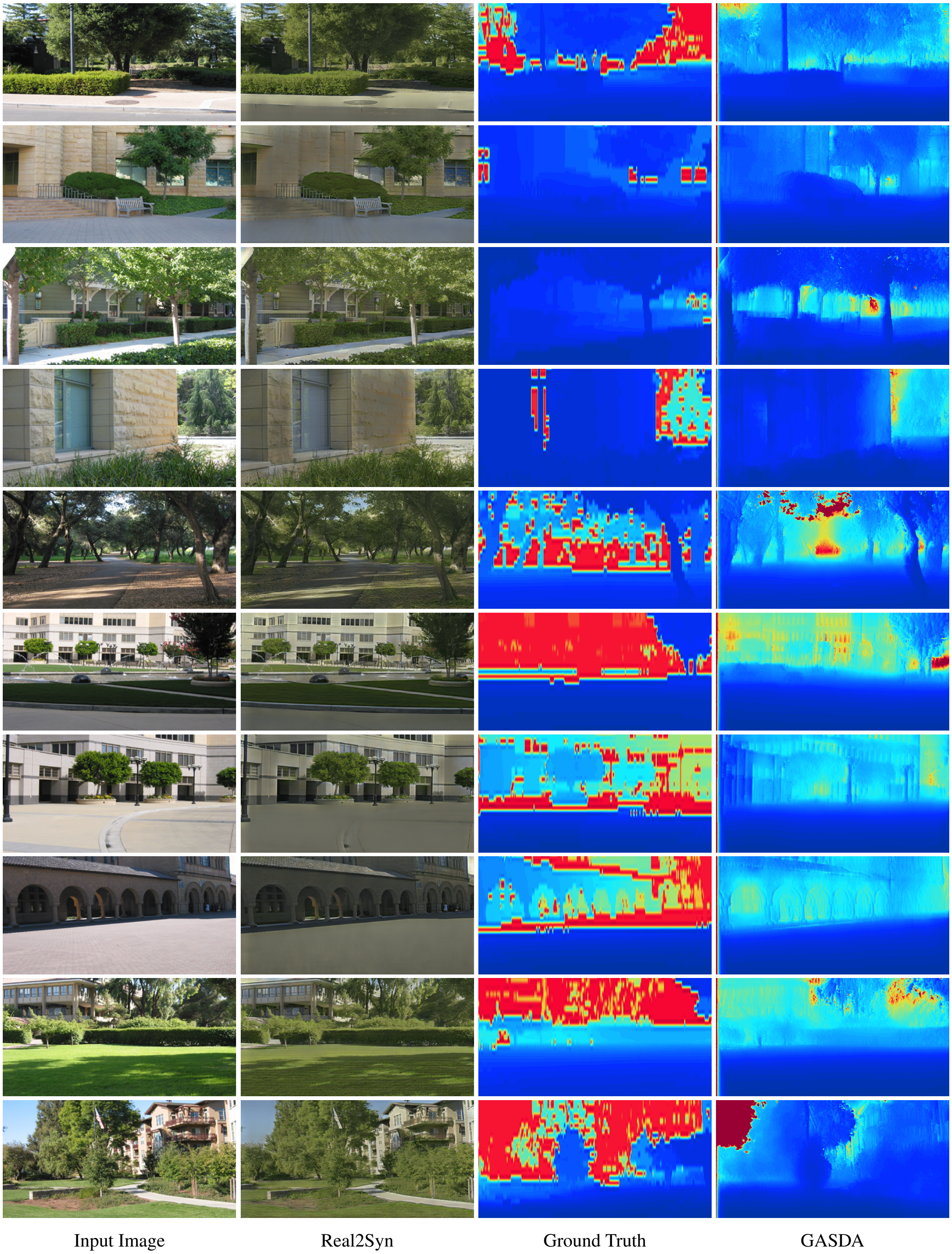
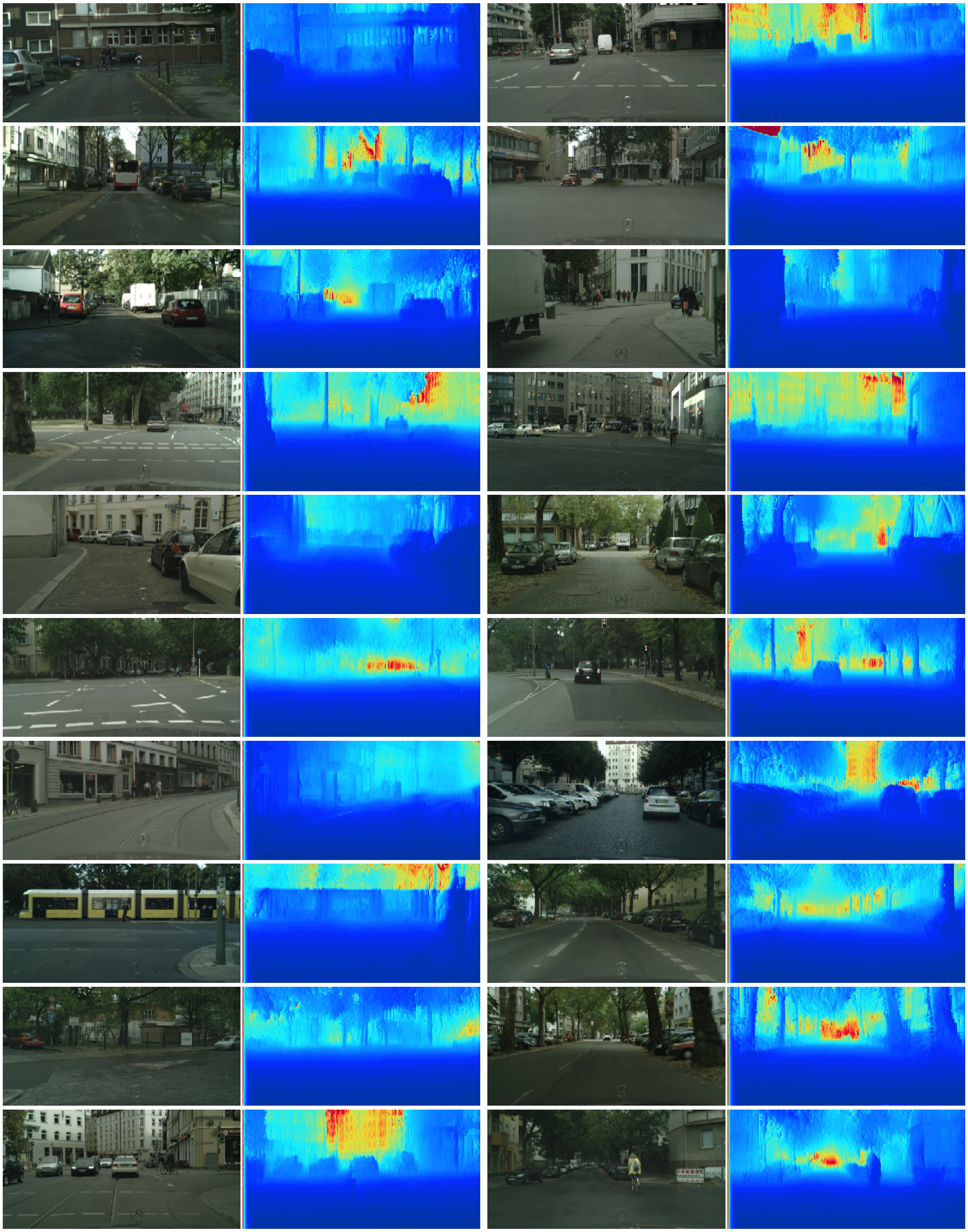
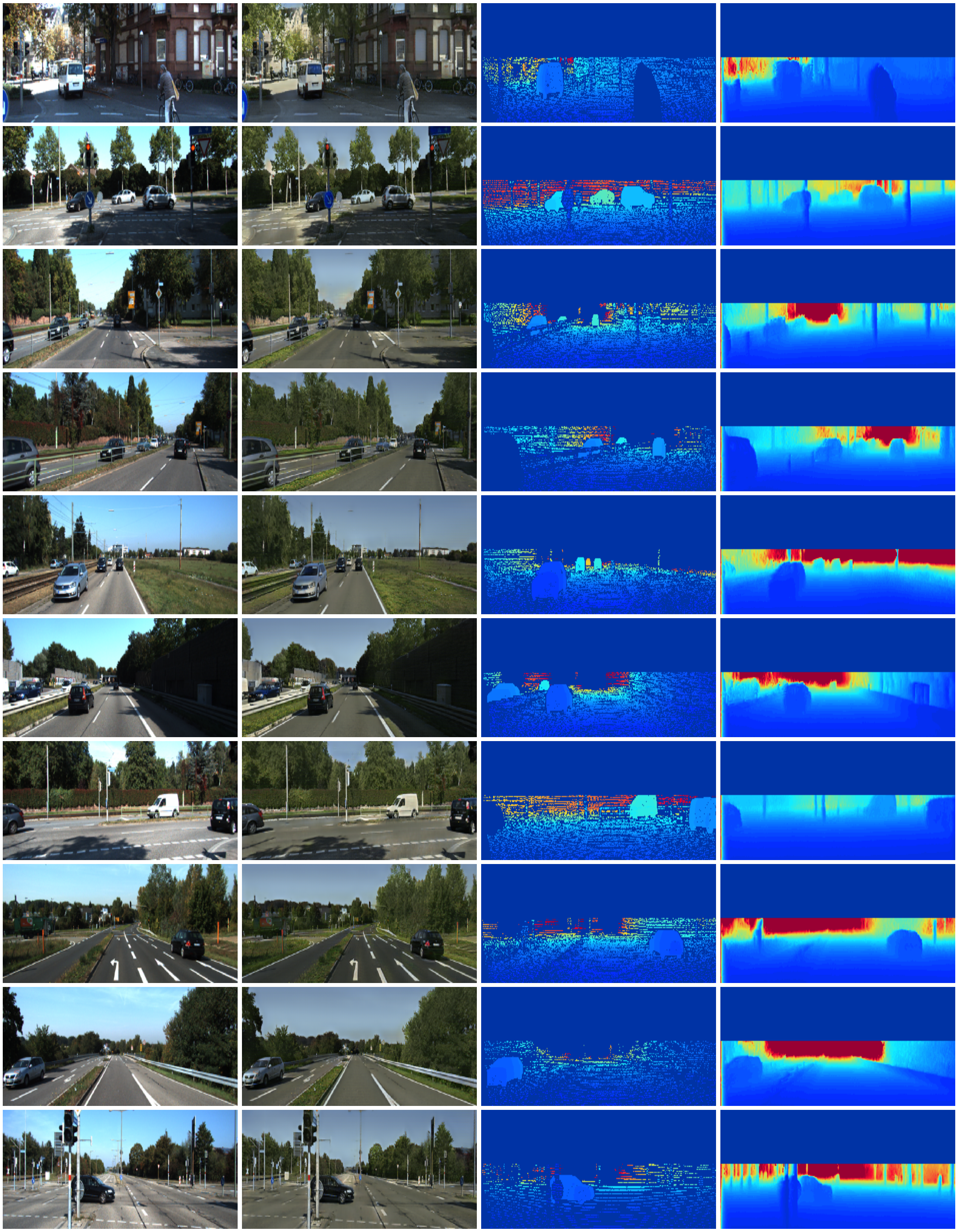


Figure 10: Qualitative results on Make3D dataset [45]. Real2Syn denotes the intermediate real-to-synthetic translation result in GASDA. The model is trained on KITTI using the split of Eigen *et al.* [9] without further fine-tuning.



Input Image GASDA Input Image GASDA

Figure 11: Qualitative results on CityScapes dataset [6]. The model is trained on KITTI using the split of Eigen *et al.* [9] without further fine-tuning.



Input Image | Real2Syn Image | Ground Truth | GASDA

Figure 12: Qualitative results on the official KITTI stereo 2015 dataset [15]. The model is trained on KITTI using the KITTI Split suggested in [16].