Estimating Image Depth in the Comics Domain

Deblina Bhattacharjee, Martin Everaert, Mathieu Salzmann, Sabine Süsstrunk
School of Computer and Communication Sciences, EPFL, Switzerland
{deblina.bhattacharjee, martin.everaert, mathieu.salzmann, sabine.susstrunk}@epfl.ch

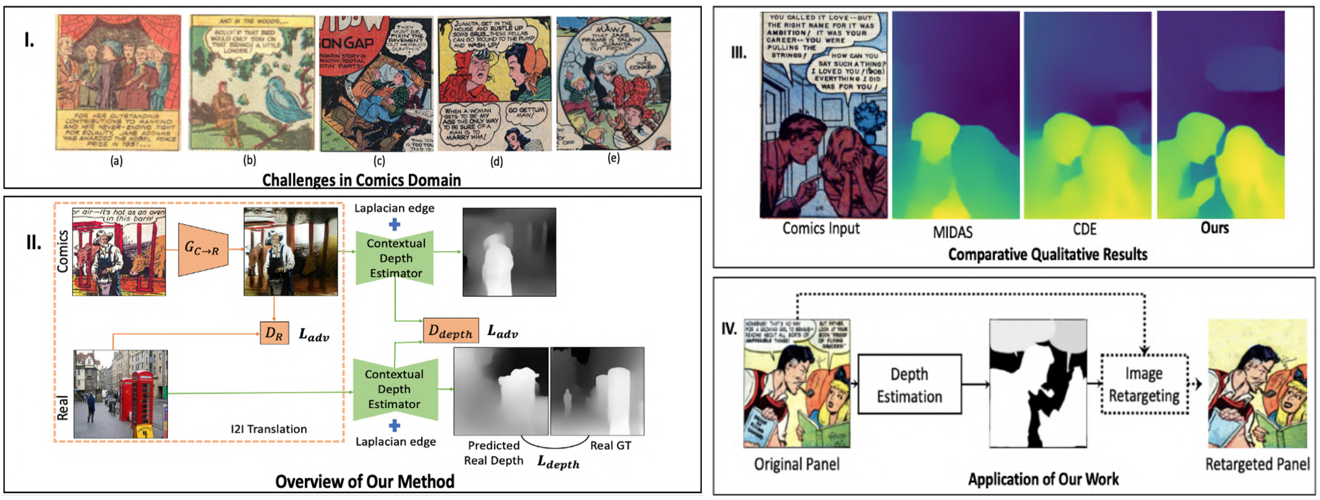

Figure 1: **I**) Estimating depth in the comics domain is subject to many **challenges**, including a) occlusions between characters; b) unusual object sizes (the bird here); c) unusual perspective; d) and e) different illustrative styles. **II**) **Overview of our model**, which uses an unsupervised I2I translation method to translate the comics image to the natural image domain and then, employs a contextual depth estimator with Laplacian edges and a feature-based GAN to ultimately predict depth. **III**) Comparative depth estimation **results** using MIDAS [34], Contextual Depth Estimation (CDE) [20] and Ours. **IV**) **Application of our method** to image retargeting, where our depth estimation model guides the retargetting model [3].

Abstract

Estimating the depth of comics images is challenging as such images a) are monocular; b) lack ground-truth depth annotations; c) differ across different artistic styles; d) are sparse and noisy. We thus, use an off-the-shelf unsupervised image to image translation method to translate the comics images to natural ones and then use an attention-guided monocular depth estimator to predict their depth. This lets us leverage the depth annotations of existing natural images to train the depth estimator. Furthermore, our model learns to distinguish between text and images in the comics panels to reduce text-based artefacts in the depth estimates. Our method consistently outperforms the existing state-of-the-art approaches across all metrics on both the DCM and eBDtheque images. Finally, we introduce a dataset to evaluate depth prediction on comics.

1. Introduction

Depth estimation for comics images can provide important information for applications such as comics image retargeting [8, 36], scene reconstruction [10] and reconfiguration of comics [35], i.e., transferring the stories from paper to an interactive graphical media, for instance, video games based on comics or comics animations. The problem of depth estimation can be framed as that of predicting a metric depth for each pixel in a given input image. In comics, the depth estimation problem is monocular, which makes it inherently ill-posed [34]. This is further complicated by the fact that most scenes in the comics domain have large content variations, object occlusions, geometric detailing (different perspectives and size scales), sparse or noisy scenes and non-homogeneous illustrations as shown in Figure 1. As a consequence, while estimating the depth of a comics scene is easy for humans, it remains highly challenging for computational models.

To address this, we explore the extensive research done in the field of monocular depth estimation over the past years, which reports computational models that leverage monocular cues, such as perspective information, object sizes, object localization, and occlusions, to estimate the depth of scenes [30]. Note that, while much work has also been done for depth estimation from stereo images [4, 40, 42] or video sequences [21, 28], such approaches do not match the monocular setting we face in the comics domain.

Because the state-of-the-art monocular depth estimation models [20, 34] have been trained on natural images, they fail to predict the depth of comics images accurately, resulting in vague, overlapping or missing objects (Figure 1). An immediate solution would be to retrain the depth models on comics images, either in a supervised manner, which would in turn require ground-truth depth annotations of comics images, or in an unsupervised manner, which would require employing domain adaptation techniques [49]. As there exist no dataset with ground-truth depth annotations for comics images and manually annotating the depth of a large number of comics images would be expensive and time-consuming, we employ an unsupervised image-to-image (I2I) translation method [5] to translate the images from the comics domain to the real one. Once translated to the real domain, we leverage the ground-truth depth of real images to train our depth model and thereby predict the depth of the translated comics⟶real image. The result of this process, compared to the direct application of a trained depth estimation network, is shown in Figure 2.

To improve the performance of the depth estimation, we exploit contextual attention, both spatially and channel-wise, as focusing on the scene context parallels how humans estimate the depth of a scene. To this end, we introduce a local context model that leverages a Laplacian edge detector to guide depth estimation. This builds on the intuition that depth features significantly depend on edge cues and yields a sharper foreground vs. background separation. Furthermore, we incorporate a feature-based GAN that encourages the inner feature representations of the depth model to follow similar distributions for the real and translated images. Additionally, we include a text detector in our model to remove the artefacts in the depth predictions arising from the text or speech balloons in comics images.

Our main contributions therefore are as follows:

- We introduce a cross-domain depth estimation method by leveraging an off-the-shelf unsupervised I2I translation method.
- We exploit the contextual information for depth prediction of a given scene. We use an inner feature-based GAN to enforce similarity between the domains, as well as a Laplacian edge detector to obtain distinct foreground vs. background separations.
- By introducing a text detector in our cross-domain depth model, we reduce the artefacts from text and speech balloons in the depth predictions, which are specific to comics.
- Finally, we introduce a benchmark dataset for comics images with 450 manually annotated image-depth pairs comprising 300 images from the standard DCM [33] validation dataset and 150 images from the standard eBDtheque [12] validation dataset. This can be used as a benchmark for future papers for depth evaluations, as there is no existing benchmark with depth annotations for comics.

Comics Input | CDE [20]- Comics Image

Translated Image | CDE [20]- Translated Image

Figure 2: **Leveraging monocular depth estimation models.** When employed directly on a comics image, the state-of-the-art monocular depth estimation model [20] fails to predict accurate depth. We therefore, translate the comics image to the natural image domain and then apply the CDE depth estimator as mentioned in [20].

Our experiments on our manually annotated benchmark show that our approach outperforms the state-of-the-art unsupervised monocular depth estimation methods across all the different comics styles.

2. Related Work

2.1. Monocular Depth Estimation

Over the past decade, there has been a significant development in monocular depth estimation. Laina et al. [25] proposed fully convolutional networks with the fast up-projection method using residual learning to model the mapping between RGB images and depth maps. Kuznietsov et al. [24] introduced a semi-supervised approach to overcome the deficiency and limitation of sparse ground-truth lidar maps. Godard et al. [11] suggested unsupervised training objective to replace the use of labeled depth maps. The network generates the left and right disparity maps and calculates the reconstruction, smoothness, and left-right consistency losses. Guo et al. [13] incorporated a synthetic depth dataset to acquire a considerable amount of ground-truth images. Subsequently, they trained a network with synthetic data and fine-tuned with a real dataset. Finally, they mitigated the domain gap between the ground-truth and synthetic dataset by distilling stereo networks. Amirkolaee and Arefi [1] constructed a depth prediction network with the encoder–decoder and skip connection structure to integrate the global and local contexts. In [20], a context based monocular depth estimation method exploits the contextual information between objects via inter object attention to extract visual cues for estimating depth. While these approaches produce improved and consistent depth results, training them is challenging because of 1) inherently different representations of depth: direct vs. inverse depth representations [14, 19], 2) scale ambiguity: for some data sources, depth is only given up to an unknown scale [6, 46, 47], 3) shift ambiguity: some datasets provide disparity only up to an unknown global disparity shift [44]. Further, in the presence of occluded regions (i.e. groups of pixels that are seen in one image but not the other), these methods produce meaningless values due to failed disparity calculations. To mitigate this, in [34], the authors propose a new loss function that is invariant to both scale and global shift so that the monocular depth estimation model can learn from diverse ground-truth depth maps obtained from disparate domains. Nevertheless, it does not generalise well to either paintings or comics domain.

With the development of image style transfer and its connection with domain adaptation, researchers adopted the style transfer and adversarial training to estimate depth maps in real scenes [2, 23], which relied on the models trained with a large amounts of synthetic data. DispNet [29] was the first network that introduced image style transfer for depth estimation. Thereafter, Zheng et al. [49] proposed a two-module domain adaptive network, T2Net, where one module was trained with synthetic and real images and reconstructed each other with the reconstruction loss and generative adversarial loss [7, 22], and these outputs were input into the other module to predict the real depth maps. As this method is close to our approach, we consider the T2Net as a baseline for comparison. Besides, there are more models with cycle consistency [48], cross-domain [13, 41], and others for domain adaptation to predict monocular depth maps. In this vein, we apply an unsupervised I2I translation method to minimize the domain disparity between comics and real world.

2.2. Domain Adaptation via I2I Translation

The advent of I2I translation methods began with the invention of conditional GAN[31], which have been applied to a multitude of tasks, such as scene translation [18] and sketch-to-photo translation [43]. While conditional GANs yield impressive results, they require paired images during training. Unfortunately, in comics⟶real I2I translation scenario, such paired training data is lacking and expensive to collect. To overcome this, cycleGAN [50], with its cycle consistency loss between the source and target domains, is a possible solution for translating the comics images to real images, thereby producing consistent images. Nevertheless, neither conditional GANs, nor cycleGAN account for the multi-modality of comics⟶real I2I translation; in general, a single comics image can be translated to real domain in many different, yet equally realistic ways. This is also due to the different artistic styles present in a single comics domain, which in turn, gives rise to intra-comics domain style variability. Addressing this issue of multi-modality, more recently, MUNIT [17] and DRIT [26] introduced solutions by learning a disentangled representation with a domain-invariant content space and a domain-specific attribute/style space. While effective, all the above-mentioned methods perform image-level translation, without considering the object instances. As such, they tend to yield less realistic results when translating complex scenes with many objects. This is also the task addressed by INIT [38] and DUNIT [5]. While INIT [38] proposed to define a style bank to translate the instances and the global image separately, DUNIT [5] proposed to unify the translation of the image and its instances, thus preserving the detailed content of object instances. We, therefore, use DUNIT [5] as our I2I translation model to translate the comics images to real domain. Once translated, we leverage a depth estimator trained with depth annotations from real images, to ultimately, predict the depth of comics images.

3. Methodology

3.1. Problem Formulation and Overview

We aim to learn a cross-domain depth mapping between two visual domains $C \subset \mathbb{R}^{H\times W\times 3}$ and $R \subset \mathbb{R}^{H\times W\times 3}$, where C is the comics domain and R is the real image domain. To this end, first we employ the DUNIT model [5] to translate the given comics image to the real domain. Second, we use a contextual monocular depth estimator on the translated image. Thus, the problem can be formulated as $D_c = f(R(C))$, where D_c is the depth prediction for the given comics image C, $R(C)$ is the comics⟶real translated image and $f(R(c))$ is the depth estimator trained on real images and applied to $R(c)$. The detailed architecture of our method is provided in Figure 3. We now explain the components of our network in more detail.

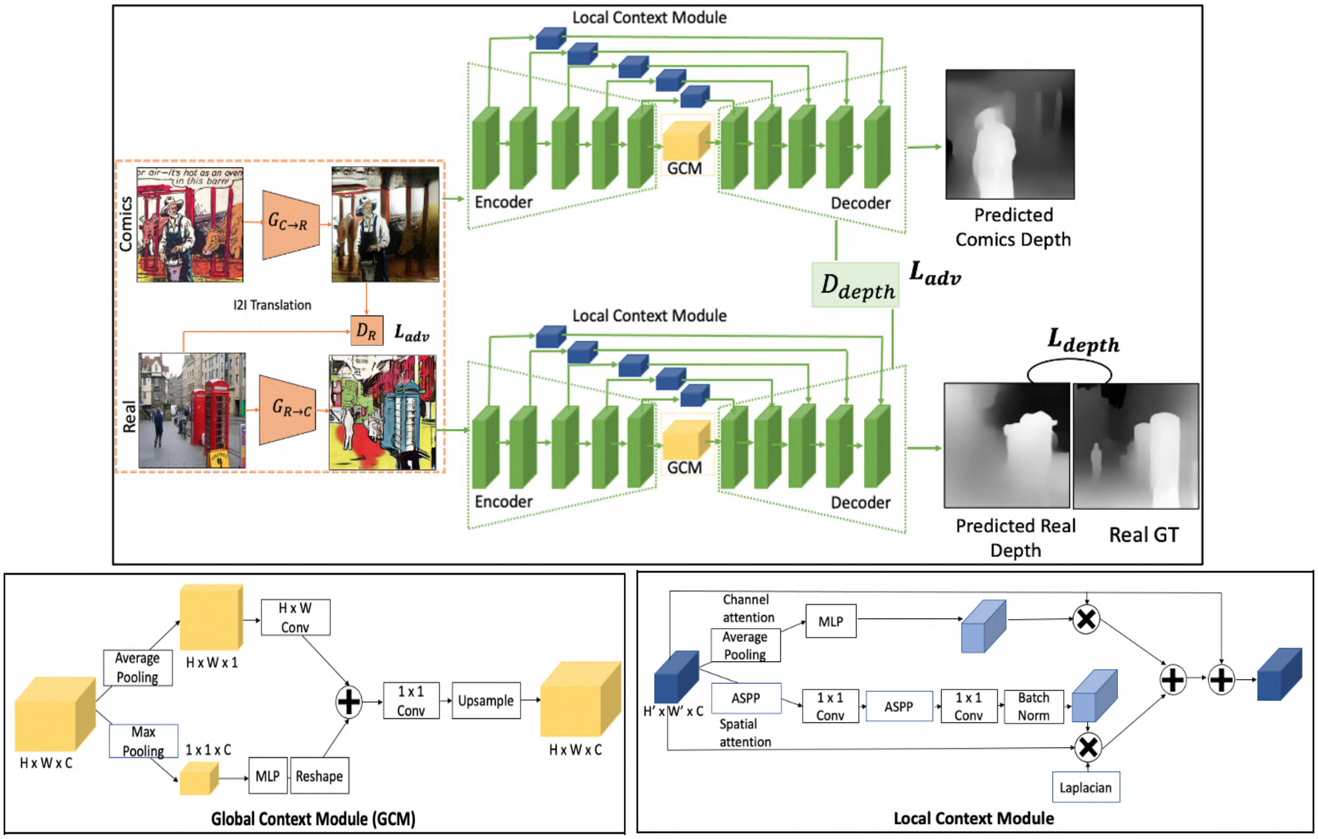

Figure 3: **Detailed overview of our architecture.** Top: Overall architecture as discussed in Section 3. Bottom Left: Global Context Module as detailed in [20]. Bottom Right: Local Context Module [20] with the added Laplacian in the spatial attention branch.

3.2. Training

To handle unpaired training images between the comics and real domains, we follow the cycle-consistency approach. In essence, this process mirrors that described in DUNIT [5]. Additionally, to study the effect of the I2I translation model on the performance of the depth estimator, we replace the DUNIT method with CycleGAN [50] and DRIT [26]. These methods do not reason about the instance-level translations and thus, perform poorly in contrast to DUNIT. We report these results in the next section. Below, we detail the loss function and training procedure for the resulting I2I translation based depth model.

Image-to-image translation module. Our method is built on the DUNIT [5] backbone which embeds the input images onto a shared style space and a domain specific content space. As such, we use the same weight-sharing strategy as DUNIT for the two style encoders (E^s_x, E^s_y) and exploit the same loss terms. They include:

- A content adversarial loss $\mathcal{L}^{content}_{adv}(E^c_x, E^c_y, D^c)$ relying on a content discriminator D^c and the two content encoders (E^c_x, E^c_y), whose goal is to distinguish the content features of both domains;
- Domain adversarial losses $\mathcal{L}^x_{adv}(E^c_y, E^s_x, G_x, D^x)$ and $\mathcal{L}^y_{adv}(E^c_x, E^{ci}_x, E^s_y, G_y, D^y)$, one for each domain, with corresponding domain classifiers D^x and D^y, corresponding domain generators G_x and G_y and instance content encoder E^{ci}_x;
- A cross-cycle consistency loss $\mathcal{L}^{cc}_1(G_x, G_y, E^c_x, E^{ci}_x, E^c_y, E^s_x, E^s_y)$ that exploits the disentangled content and style representations for cyclic reconstruction [45];
- Self-reconstruction losses $\mathcal{L}^x_{rec}(E^c_x, E^{ci}_x, E^s_x, G_x)$ and $\mathcal{L}^y_{rec}(E^c_y, E^s_y, G_y)$, one for each domain, ensuring that the generators can reconstruct samples from their own domain;
- KL losses for each domain $\mathcal{L}^x_{KL}(E^s_x)$ and $\mathcal{L}^y_{KL}(E^s_y)$ encouraging the distribution of the style representations to be close to a standard normal distribution;
- Latent regression losses $\mathcal{L}^x_{lat}(E^c_x, E^{ci}_x, E^s_x, G_x)$ and $\mathcal{L}^y_{lat}(E^c_y, E^s_y, G_y)$, one for each domain, encouraging

the mappings between the latent style representation and the image to be invertible;

- An instance consistency loss $\mathcal{L}_1^{ic}(P_{tl}^{xi}, P_{tl}^{yi}, P_{br}^{xi}, P_{br}^{y})$ encouraging the same object instances to be detected in the source domain image and in the corresponding image after translation, where $P_{(.)}^{(.)}$ are the bounding box top-left and bottom-right corner pixels for detected instances in the two domains.

During training, the I2I module is trained along with the depth estimation module in an end-to-end manner. It has been observed in [2, 49] that an end-to-end approach yields consistent results on unknown domains, though it comes with a computational overhead. In our method, this computational cost depends mainly on the employed I2I translation module. For instance, DUNIT [5] has a greater computational overhead than DRIT [26] or CycleGAN [50]. For further details we point the reader to the supplementary material.

Depth estimation module. As shown in Figure 3, our depth estimation module is an encoder-decoder model with skip connections in between the encoder and decoder layers. These skip connections model the local context in between the visual features by taking into account the spatial and channel attention. The architecture of our depth estimator is inspired from [20]. It relies on a Global Context Module (GCM), mirroring that of [20], which explores the context of the entire scene, whereby it computes the spatial and channel attention between the objects present in the global image. To this end, the GCM is placed at the end of the encoder to obtain the global context information and pass meaningful features to the decoder. We further complement the GCM with a local context module processing the features extracted at different layers in the encoder of the depth estimator as shown in [20]. Moreover, to clearly contrast the edge boundaries of the objects, we incorporate a Laplacian edge detector [16] to the spatial branch of the local context module. Since depth leverages low-level visual cues, such as edge information, we have observed this Laplacian to facilitate depth estimation. In particular, the local context module feature (shown in blue in Figure 3), extracted by the encoder of the depth estimator, is processed spatially and channel-wise before being fed into the decoder layer. While the channel-wise processing mirrors that of CDE [20], the spatial processing (or the spatial attention branch as shown in Figure 3) employs multiple ASPP [15] and convolutions to obtain a spatially-pooled feature, which is then multiplied with the original local context module feature and the Laplacian [16]. Finally, the features from both the spatial and channel branch are added to the original feature, to produce the processed local context feature. This feature is fed into the decoder layer.

Our method uses two depth estimators, one taking the real images as input and the other the translated images. We use the zero-shot cross domain MIDAS model [34] to generate pseudo ground-truth depth for the real domain. Note, however, that we could use any existing real-image dataset with ground-truth depth annotations, such as KITTI [9] or NYU [32]. However, these datasets are restrictive on the diversity of their scenes, i.e., they are not representative of the extreme scene diversity in comics that contain both indoor and outdoor scenes. Therefore, we use the MIDAS model, which was trained on a collection of five diverse real-world datasets comprising both indoor and outdoor scenes. We generate the pseudo ground-truth only once, before training our depth estimators.

Nevertheless, MIDAS fails when directly applied on comics images (shown in supplementary Figure 2), hence the need for our cross-domain context aware depth estimators. To train them, we initialize both with the MIDAS weights, setting a low learning rate of 10^{-6} to update the weights for 100 epochs with the Adam optimizer and the default hyper-parameters of [20]. During the training phase, we use a shift and scale-invariant log loss function [34] as objective function L_{depth} for the depth estimator in the real domain. It can be expressed as

$$L_{depth}(y, y^*) = \frac{1}{n}\sum_i d_i^2 - \frac{1}{2n^2}\left(\sum_i {d_i}^2\right) , \quad (1)$$

where $d_i = log(y_i) - log(y_i^*)$, y is the predicted depth, y^* is the pseudo ground-truth depth in the real domain and n is the number of pixels indexed by i.

As it learns the depth mappings, the depth estimator in the real domain shares its weight with the estimator in the comics$\longrightarrow$real translated domain. We then add an adversarial loss L_{adv} to train the feature-based GAN between the two depth estimators [49], which encourages the inner feature representations of the two depth estimators to share similar distributions, since both stylistically represent real images. This loss is written as

$$\begin{aligned} L_{adv}(f, D_{depth}) = E_{f_{c'} \in f_C}[log(1 - D_{depth}(f_{c'}))] \\ +E_{f_{r'} \in f_R}[log(D_{depth}(f_{r'}))] , \end{aligned} \quad (2)$$

where f_C and f_R represent the encoded features extracted by the encoder of the depth estimators in the translated domain and the real domain respectively, and D_{depth} is the discriminator of the feature GAN.

Altogether, we write the overall objective function to train our depth estimators as

$$\begin{aligned} L_{obj}(f, D_{depth}) = \alpha_{adv} L_{adv}(f, D_{depth}) + \\ \alpha_{depth} L_{depth}(f) . \end{aligned} \quad (3)$$

Text detection module. When the comics images are translated to the real domain, the translated images comprise text areas or speech balloons, which are in turn unknown to the depth estimator trained on the real domain. This leads to text-based artefacts in the depth results as the depth estimator considers such text areas as objects. Therefore, to control the position of the text areas in the translated images, we train a U-net [37] in a supervised manner using the eBDtheque dataset [12], which contains text/speech balloon annotations. We mask the depth maps by multiplying them pixel-wise with the compliment of the text-area mask, before using the L1-loss between the (masked) pseudo ground-truth depth and the depth predictions. The detailed architecture for our method with the text detection module is given in the supplementary material.

4. Experiments and Results

To validate our method, we conduct experiments on the following datasets.

4.1. Datasets

The main datasets used for this work are DCM [33] and eBDtheque [12] for the comics domain and the COCO dataset [27] for the real-world domain. The **DCM** dataset comprises 772 full-page images with multiple comics panel images within. We extract 4470 single panel images from these full-page images using the panel annotations. Note that the panel annotations do not contain depth information. We thus, use these DCM panel images to train the I2I model. The **eBDtheque** dataset contains 100 full-page images with multiple comics panel images within. Again, we extract 850 single panel images as before. The eBDtheque dataset contains annotations for speech balloons and text lines, which we use to train a U-net [37] to predict the text areas in a comics image. The detected text areas are then used by our depth model to remove text-based artefacts from the depth predictions. We employ the **MS-COCO** dataset [27], comprising 5000 real-world images, as real-world domain to train the I2I model.

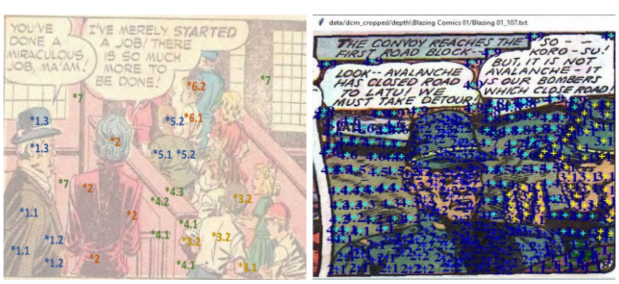

Figure 4: **Benchmark for evaluation.** Left: Illustration of the idea of inter-object and intra-object depth ordering, used to annotate the comics images. The closer object is assigned a lower first number l_1; and the closer point within the same object is assigned a lower second number l_2. Right: Annotated example from the benchmark.

Benchmark for evaluation. To evaluate and compare the different depth models, we introduce a benchmark including 300 DCM [33] images and 150 eBDetheque [12] images, from their validation set, along with the corresponding manually annotated ground-truth depth orderings, as illustrated in Figure 4. To manually annotate their depth, we carefully select 450 images from DCM and eBDtheque validation sets, such that, they contain diverse scenes across ten different artistic styles. These images were further tested for inter-observer variability, for instance, their diversity and artistic styles were analysed by three comics domain experts. Further, all three observers tested the manually annotated depth of comics images. We use depth orderings to annotate the images. In particular, the image pixel coordinates (x, y) are assigned two numbers (l_1, l_2). The first number, l_1, represents the inter-object depth ordering, such that two different l_1 values imply two different objects. Closer objects are assigned a lower number. The second number, l_2, represents the intra-object depth ordering, such that annotations with the same first number but different second numbers indicate that the two points belong to the same object. A lower l_2 value indicates a closer point on the same object.

4.2. Evaluation Metrics

To evaluate our method, we evaluate the following four standard performance metrics, as used in [20, 34].

Absolute relative difference (AbsRel). The absolute relative difference is given by $\frac{1}{|N|}\sum_{y\in N}|y-y^*|/y^*$ where N is the number of available pixels in the manually annotated ground-truth.

Squared relative difference (SqRel). The squared relative difference is defined as $\frac{1}{|N|}\sum_{y\in N}\|y-y^*\|^2/y^*$.

Root mean squared error (RMSE). The root mean squared error is defined as $\sqrt{\frac{1}{|N|}\sum_{y\in N}\|y-y^*\|^2}$.

RMSE (log). The RMSE (log) is defined as $\sqrt{\frac{1}{|N|}\sum_{y\in N}\|\log y-\log y^*\|^2}$.

4.3. Quantitative Results

To evaluate our method, we compare it with the following four state-of-the-art depth estimation approaches.

- T2Net [49], which comprises a depth prediction model trained on synthetic image-depth pairs.
- Song et al. [39], which incorporates a Laplacian pyramid into the decoder architecture. In particular, the encoded features are fed into different streams for decoding depth residuals, defined by the Laplacian pyramid, and the corresponding outputs are progressively combined to reconstruct the final depth map from coarse to fine scales.

- MIDAS [34], which introduces a scale and shift-invariant loss to estimate depth from a large collection of mixed real-world datasets, thereby presenting a depth model that generalises across multiple real-world datasets.

- CDE [20], which proposes an architecture that leverages contextual information in a given scene for monocular depth estimation. Thus, using the contextual attention it obtains meaningful semantic features to enhance the performance of the depth model.

We report the standard evaluation metrics for our method in comparison with the four state-of-the-art methods in Table 1 and Table 2 on the DCM and eBDtheque images, respectively. Note that to report the performance metrics, we compare the predicted depth by each method with our manually annotated ground-truth depth. For the results in Table 1, we use the 300 manually annotated DCM image-depth pairs from our benchmark. Further, for the results in Table 2, we use the 150 manually annotated eBDtheque image-depth pairs from our benchmark. Our method outperforms the baselines on all the performance metrics for both DCM and eBDtheque images. Note that to evaluate the performance of the four state-of-the-art methods, the comics image is translated to the real domain using a pretrained DUNIT model and then, the respective methods are applied to predict its depth. This is imperative as the above state-of-the-art methods are trained on real domain, and thus to evaluate them fairly on comics, we translate the comics image to the real domain. To maintain consistency, we also evaluate our approach on the translated comics$\longrightarrow$real image. Nevertheless, our approach can also be directly applied on a comics image to predict its depth. We show this qualitatively in the supplementary material.

Method	AbsRel↓	SqRel↓	RMSE↓	RMSE log↓
T2Net [49]	0.351	0.416	1.117	0.415
Song et.al. [39]	0.339	0.401	1.098	0.402
MIDAS [34]	0.309	0.381	1.033	0.375
CDE [20]	<u>0.304</u>	<u>0.374</u>	<u>1.024</u>	<u>0.367</u>
Ours	**0.251**	**0.318**	**0.971**	**0.305**

Table 1: **Quantitative comparison (DCM images).** We compare our approach with the state-of-the-art methods on the translated DCM validation images [33] from our benchmark. We report the Absolute Relative Difference (AbsRel), Squared Relative Difference (SqRel), Root Mean Squared Error (RMSE), and RMSE log (lower the better). Our contextual depth estimator with the feature-based GAN, Laplacian and text detection module gives the best result. The best results are in bold and the second-best are underlined.

Method	AbsRel↓	SqRel↓	RMSE↓	RMSE log↓
T2Net [49]	0.491	0.555	1.459	0.777
Song et.al. [39]	0.479	0.520	1.431	0.711
MIDAS [34]	<u>0.419</u>	<u>0.503</u>	1.416	0.659
CDE [20]	0.424	0.511	<u>1.415</u>	<u>0.647</u>
Ours	**0.376**	**0.448**	**1.364**	**0.553**

Table 2: **Quantitative comparison (eBDtheque images).** We compare our approach with the state-of-the-art methods on the translated eBDtheque validation images [12] from our benchmark. We report the Absolute Relative Difference (AbsRel), Squared Relative Difference (SqRel), Root Mean Squared Error (RMSE), and RMSE log (lower the better). Our contextual depth estimator with the feature-based GAN, Laplacian and text detection module gives the best result. The best results are in bold and the second-best are underlined.

4.4. Qualitative Results

In Figure 5, we compare our method with the depth predictions obtained by MIDAS [34] and CDE [20]. The examples demonstrate that our network can benefit from I2I translation in addition to the feature-based GAN and Laplacian. Moreover, we also qualitatively show the effect of our text-detection module. For instance, in the middle row of Figure 5, while MIDAS and CDE have text-based artefacts in the predictions, including vague depth values in the background from the speech balloons and incorrect depth from the text box in the foreground, our method correctly removes the speech balloon artefacts. Further, our model predicts the human object in the same depth plane as that of the text box in the foreground. Note that these predictions were verified by comics domain experts. Our method threfore, yields sharper depth maps with clearer foreground vs. background separation and with well-defined object edges. Furthermore, in contrast to the baselines, the depth predictions by our method show greater consistency in their intra-object and inter-object depth values.

4.5. Ablation Study

We now evaluate different aspects of our method. First, we study the influence of the I2I translation module on our depth model (including the feature GAN, Laplacian and the text module). To this end, we compare the results obtained using our depth model with the different state-of-the-art I2I method, namely, cycleGAN [50], DRIT [26] and DUNIT [5]. We report the AbsRel, SqRel, RMSE and RMSE (log) on the DCM validation images [33] from our benchmark in Table 3. We observe that DUNIT consistently improves the results across all metrics, thereby demonstrating the benefits of instance-level translations on our method, in contrast to the image-level translations in cycleGAN and DRIT.

Comics input MIDAS [34] CDE [20] Our

Figure 5: **Qualitative comparison of depth estimation** on the translated DCM validation images [33] from our benchmark, using the text detection module (top row and middle row) and without using the text detection module (bottom row). We show, from left to right, the input image in the comics domain, the result using the MIDAS [34] model directly on the translated comics image, the result using the CDE [20] model directly on the translated comics image, and **Our** model applied to the translated comics image, respectively.

Method	AbsRel↓	SqRel↓	RMSE↓	RMSE log↓
CycleGAN [50]	0.282	0.346	0.995	0.329
DRIT [26]	<u>0.269</u>	<u>0.333</u>	<u>0.983</u>	<u>0.317</u>
DUNIT [5]	**0.251**	**0.318**	**0.971**	**0.305**

Table 3: **Ablation Study on the effect of I2I model.** We compare the effect of the different I2I translation model on our method. We report the four standard performance metrics (lower the better). Our method with the DUNIT model gives the best result. The best results are in bold and the second-best are underlined.

We then turn to exploring the effect of the feature GAN, Laplacian and text detection module on our method. To this end, we add each of these components one-by-one to the baseline approach comprising the DUNIT model and the CDE model, shown as I2I+depth in Table 4. Note that this baseline approach is trained in an end-to-end manner. We report the standard four performance metrics on the DCM [33] images from our benchmark in Table 4. We show that the end-to-end baseline approach outperforms the CDE [20] method when applied directly to the translated comics images, as shown in Table 1. This solidifies the benefits of an end-to-end training approach. Moreover, the addition of each component of our method consistently improves the performance across all metrics. All the images were kept constant for the study of all the network components. We show qualitative results from this ablation study in our supplementary material.

Method	AbsRel↓	SqRel↓	RMSE↓	RMSE log↓
I2I + Depth	0.301	0.369	1.022	0.362
Feature GAN	0.270	0.339	0.994	0.322
Laplacian	<u>0.257</u>	<u>0.322</u>	<u>0.976</u>	<u>0.313</u>
Text Module	**0.251**	**0.318**	**0.971**	**0.305**

Table 4: **Ablation Study on the effect of the different network components.** We compare the effect of the different network components, namely, the feature GAN, Laplacian and text module on our method. We report the four standard performance metrics (lower the better). The above network components are added one-by-one and we observe that *our model with feature GAN, Laplacian and text module* outperforms on all performance metrics. The best results are in bold and the second-best are underlined.

5. Conclusion

We have introduced an approach to estimate image depth in the comics domain using unsupervised I2I translation to adapt the comics images to the real domain. To this end, we have leveraged a modified context-based depth model trained on real-world images with Laplacian. We also, have added a feature GAN approach to the depth estimators to enforce the semantic similarity between the translated and real images. We have further added a text-detection module to remove text-based artefacts in the depth predictions. To validate our experiments, we introduce a benchmark with manually annotated depth for images from the validation set of DCM and eBDtheque datasets, as there is no existing benchmark with depth annotations. In our experiments, our I2I translation-based modified depth estimators with Laplacian, feature GAN and text-detections, outperform the state-of-the-art methods. This is the first automated method to predict depth for comics images. Therefore, this work can be used for applications like comics image retargeting, scene reconstruction, comics animations or repurposing comics to augmented reality.

Acknowledgement. This work was supported in part by the Swiss National Science Foundation via the Sinergia grant CRSII5−180359.

References

[1] Hamed Amini Amirkolaee and Hossein Arefi. Monocular depth estimation with geometrical guidance using a multi-level convolutional neural network. *Applied Soft Computing*, 84:105714, 2019.

[2] Amir Atapour-Abarghouei and Toby P. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2800–2810, 2018.

[3] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3):10–es, July 2007.

[4] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3d: Stereo depth estimation via binary classifications. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1597–1605. IEEE Computer Society, 2020.

[5] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018.

[7] Yuan Chen, Yang Zhao, Wei Jia, Li Cao, and Xiaoping Liu. Adversarial-learning-based image-to-image transformation: A survey. *Neurocomputing*, 411:468–486, 2020.

[8] Xiaoting Fan, Jianjun Lei, Jie Liang, Yuming Fang, Nam Ling, and Qingming Huang. Stereoscopic image retargeting based on deep convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021.

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

[10] Margrit Gelautz, Michael Bleyer, Danijela Markovic, and Christoph Rhemann. 3d scene reconstruction by stereo methods for analysis and visualization of sports scenes. In Arnold Baca, Martin Lames, Keith Lyons, Bernhard Nebel, and Josef Wiemeyer, editors, *Computer Science in Sport - Mission and Methods*, number 08372 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2008. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Cl'ement Gu'erin, Christophe Rigaud, Antoine Mercier, Farid Ammar-Boudjelal, Karell Bertet, Alain Bouju, Jean-Christophe Burie, George Louis, Jean-Marc Ogier, and Arnaud Revel. ebdtheque: a representative database of comics. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1145–1149, 2013.

[13] 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.

[14] Kin Gwn Lore, Kishore Reddy, Michael Giering, and Edgar A. Bernal. Generative adversarial networks for depth map estimation from rgb video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Lecture Notes in Computer Science*, page 346–361, 2014.

[16] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2006.

[17] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *European Conference on Computer Vision (ECCV)*, abs/1804.04732, 2018.

[18] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36:1–14, 07 2017.

[19] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[20] Doyeon Kim, Sihaeng Lee, Janghyeon Lee, and Junmo Kim. Leveraging contextual information for monocular depth estimation. *IEEE Access*, 8:147808–147817, 2020.

[21] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[22] Aran CS Kumar, Suchendra M. Bhandarkar, and Mukta Prasad. Monocular depth prediction using generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 413–4138, 2018.

[23] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R. Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[24] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[25] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks, 2016.

[26] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision (ECCV)*, 2018.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[28] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 39(4), 2020.

[29] Nikolaus Mayer, Eddy Ilg, Philip Häusser, 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] Alican Mertan, Damien Jade Duff, and Gozde Unal. Single image depth estimation: An overview, 2021.

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[32] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[33] Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Digital comics image indexing based on deep learning. *Journal of Imaging*, 4(7), 2018.

[34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[35] Andreas Rauscher, Daniel Stein, Jan-Noël Thon, and Park. High-quality depth from uncalibrated small motion clip. In *Comics and Videogames: From Hybrid Medialities to Transmedia Expansions (1st edition)*, 2020.

[36] Christophe Rigaud. Segmentation and indexation of complex objects in comic book images. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 14, 12 2014.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[38] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S. Huang. Towards instance-level image-to-image translation. *CoRR*, abs/1905.01744, 2019.

[39] Minsoo Song, Seokjae Lim, and Wonjun Kim. Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021.

[40] Vladimir Tankovich, Christian Häne, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching, 2021.

[41] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised domain adaptation for depth prediction from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2396–2409, 2020.

[42] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo, 2019.

[43] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*, pages 364–374. Springer, 2013.

[44] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. *CoRR*, abs/1904.11112, 2019.

[45] Chengjia Wang, Gillian Macnaught, Giorgos Papanastasiou, Tom MacGillivray, and David E. Newby. Unsupervised learning for cross-domain medical image synthesis using deformation invariant cycle consistency networks. *CoRR*, abs/1808.03944, 2018.

[46] Delong Yang, Xunyu Zhong, Dongbing Gu, Xiafu Peng, and Huosheng Hu. Unsupervised framework for depth estimation and camera motion prediction from video. *Neurocomputing*, 385:169–185, 2020.

[47] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency, 2017.

[48] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9788–9798, 2019.

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

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