Estimating Image Depth in the Comics Domain (Supplementary)

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

In this supplementary material, we provide details about the text-detection module, additional qualitative comparison for the state-of-the-art methods with our approach, qualitative results for the ablation study of our network and an analysis on the computational cost of our network components. The document is structured as follows:

- Section 1: Text-detection Module
- Section 2: Qualitative Comparison- Depth Results
- Section 3: Qualitative Results- Ablation Study
- Section 4: Computational Cost Analysis

1. Text-detection Module

The generated real-images from the DUNIT [2] model have speech-balloons or text present in them, which are not recognised by the depth estimators trained on real-domain images. Therefore, the predicted depths contain text-based artefacts. In order to remove these artefacts, we use the text-detection module shown in Figure 1. Our text-detection module is a U-Net [9], trained in a supervised manner, on the text/ speech-balloon annotations from the eBDtheque [3] dataset. The trained U-Net [9] is then, used on the DCM [7] training images to detect the text/ speech-balloon areas in them, in the form of text masks. These text masks are then used to generate the text adder 'ground-truth' given by $(1-M)A+MB$, where M is the text mask, $(1-M)$ is its complement, A is the comics⟶real translated image (but without the text area) and B is the original comics image (containing the text area). Once the text adder 'ground-truth' is created, we train a text-adder generator with A, B and M as input. This generator takes the position of the mask, M, in the original comics image, B, and applies this positional information onto the comics⟶real translated image, A, to create a well-defined text area on the translated image. This generated output is trained using an $L1$ loss with the text adder 'ground-truth'. The reason to create a translated image with a well-defined text area is shown in Figure 1, top row, where we can see that a translated image when generated without the text area information contains text-based artefacts, which in turn, gives incorrect depth values after being fed into the depth estimator. However, the text-adder generator output produces no such text-based artefacts and gives a better depth prediction.

After the translated image with a well-defined text area is created, its fed into our depth estimator to predict the depth of the translated image with the text. Concurrently, the real image is passed to the other depth estimator to predict the depth of the real image. Both these estimators are trained in an end-to-end manner. Furthermore, to predict the depth of the translated image without the depth values from the text masks, we multiply the complement of the text mask with the prediction. This results in a clean depth prediction without any text-based artefacts. During inference, our approach can be directly applied on the original comics image with text. However, for fair comparison with the baseline approaches, we translate the comics image with text to a real image using a pretrained DUNIT (without the text-detection module) and then apply the different methods to predict their depth. As our approach has been trained with text information, it learns to separate the text-based artefacts and thus, produces a superior depth map. In Figure 3, last column, we observe the effect of our text module on the depth predictions for an input comics image from the DCM validation set of our benchmark (Please zoom in to observe the differences in the depth predictions).

Note that for our final approach (consisting I2I, depth, feature GAN, Laplacian and the text module), we use comics images without text areas to train our I2I module. This is done to facilitate the generation of real images without text artefacts (referred to as 'A' in Figure 1). To this end, we discuss the method to generate the original comics images without the text areas, in what follows.

Generating the comics-without-text dataset. To remove the text areas from the original comics, we randomly crop the original images along with their respective text mask prediction obtained by the trained U-Net [9], to a 384 x 384 size. We then, decrease the crop size by 1 unit per dimension, i.e., the image is cropped to 383 x 383, followed by 382 x 382, and so on. We repeat this process until the maximum area of the text in the image is 3% of the total image. After cropping, these images were checked manually for any remaining text areas and we found that none of the images contained significant text in them.

2. Qualitative Comparison- Depth Results

We show the depth predictions on the translated comics images as reported in Figure 5 of the main paper. Further, we show that the state-of-the-art methods like MIDAS [8] and CDE [5], which are trained on real-world images, fail to predict depth accurately when applied to comics images directly. Specifically, as seen in Figure 2, MIDAS is unable to predict the depth of the sample DCM [7] validation image from our benchmark, though it is trained on a large collection of real-world images from five different real-world datasets. This raises the need for applying these methods on a comics⟶real translated image. As seen in Figure 2, the baseline methods of MIDAS and CDE (trained on real images), benefit from the I2I translations. During inference, we first, translate the original comics image to the real domain by using a pretrained DUNIT [2] model and then, we apply the baseline depth estimators on these translated images. Nevertheless, our approach can predict the depth on both the translated image and the original comics image, while outperforming all the baselines in both the scenarios. This is because our approach is trained in an end-to-end manner along with the I2I module. Note that, we could also train all the baseline models (including T2Net [11], Song et.al [10] and MIDAS [8]), from scratch, in an end-to-end manner with our I2I model, but this was not addressed. This is because we observed that training the baseline method of CDE [5] from scratch, in an end-to-end manner, results in poorer results than our approach, as shown in Table 4 (first row) in our main paper. Note that though CDE was the best performing baseline method, it fails in comparison to our approach.

3. Qualitative Results- Ablation Study

We validate the results observed in Table 4 of the main paper with additional qualitative results. In Figure 3, we observe the effect of each of our network components on the depth predictions when applied to a translated comics image from the DCM validation set of our benchmark. Note that each network component was added one-by-one. We see, qualitatively, that the DUNIT (I2I)+ CDE (D) method, when trained in an end-to-end manner, outperforms the baseline CDE [5] method (cross-referring to Figure 2- first row and third column). We also see that the addition of the feature-based GAN (FG) greatly benefits the depth predictions as it encourages the similarity in distribution between the comics and the real domain. Moreover, the Laplacian (L) when added to our depth estimator, refines the edge contrasts and gives a better depth prediction. However, some text-based artefacts still remain in the depth prediction, resulting in vague depth values. To remedy this, we add the text module (TM) to finally, have superior depth predictions as seen in Figure 3 here and Table 4 in our main paper.

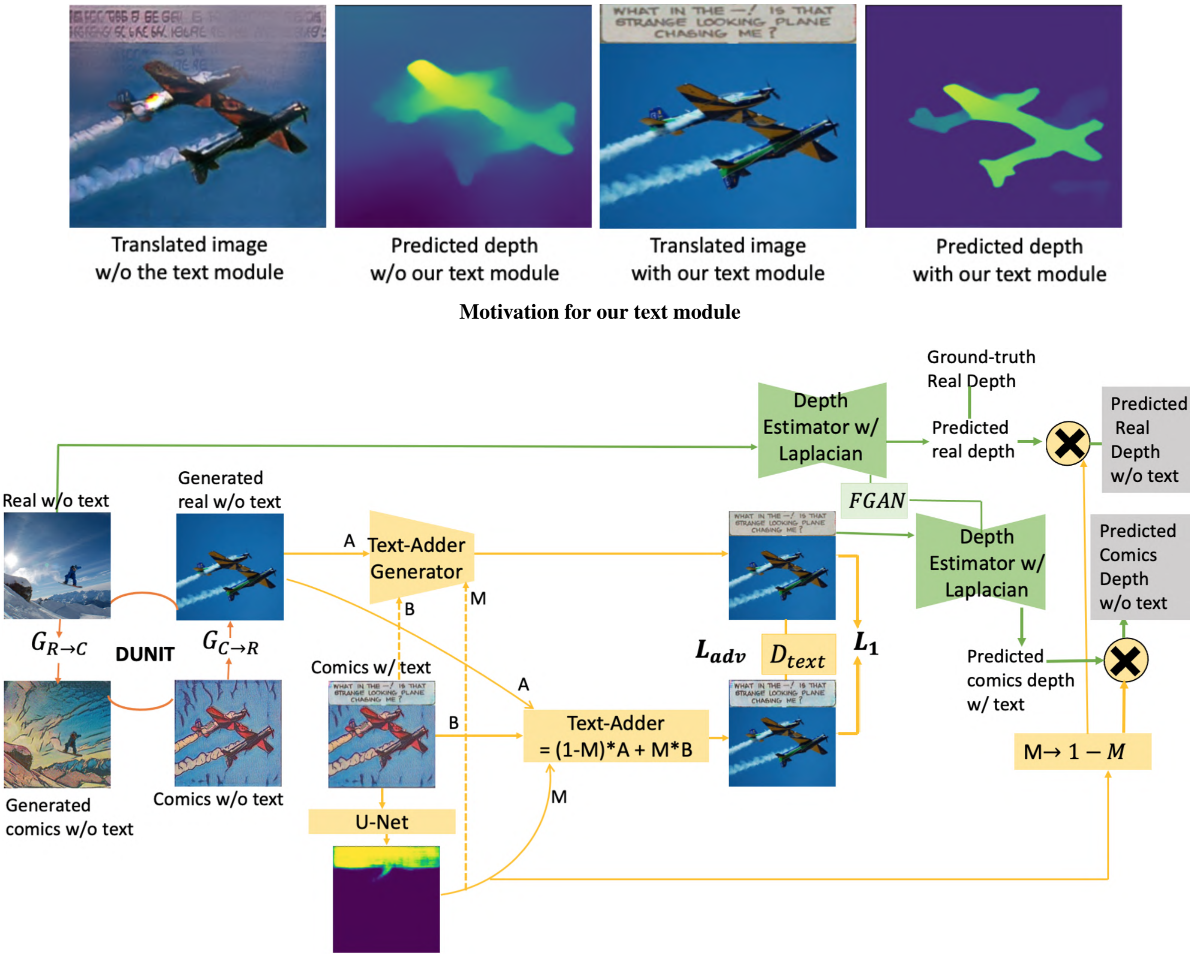
Method	w/ DUNIT [2]	w/ DRIT [6]
I2I	66%	60%
Depth (D)	17%	17%
Feature GAN (FG)	4%	10.33%
Laplacian (L)	1%	1.33%
Text Module (TM)	12%	11.34%
Total	**36h**	**27h**

Table 1: **Computational cost of training the different network components.** We compare the cost of the different network components, namely, the I2I, depth, feature GAN, Laplacian and text module in our method. We report the percentage of the total computational time taken by each of these components. We report that the I2I module dominates the training time. Note that the above methods were trained using 4 GPUs following consistent resolution for all the input images and constant batch size.

4. Computational Cost Analysis

We have seen, thus far, that training our approach in an end-to-end manner improves the predicted depth maps and thereby benefits our method. However, this leads to a computational overhead. We report the computational cost incurred by the different components of our network, when trained in an end-to-end manner in Table 1. We see that the training of the I2I module dominates the computational time, regardless of the I2I method employed. This was followed by the training of the depth estimators. Note that the training in both the scenarios (i.e. using DUNIT and DRIT as the I2I module) was done using 4 V100, 7 Tflops GPUs with 32 GB memory. The total time taken by our approach with DUNIT [2] is 36 hours, while that with DRIT [6] is 27 hours. The extra computational time for DUNIT comes from the instance-level translations. Nevertheless, the inference time for both the methods are comparable and is equal to 217 milliseconds for the method employing DUNIT and 203 milliseconds for the one with DRIT, processed on a single V100 GPU.

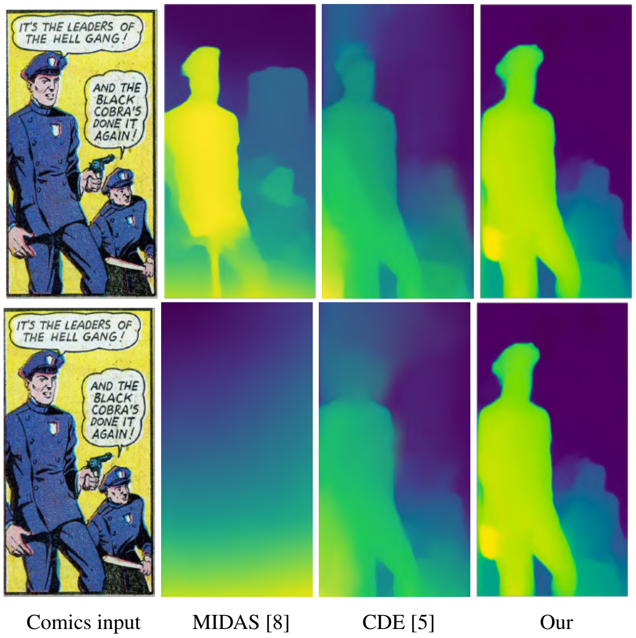
Acknowledgement. This work was supported in part by the Swiss National Science Foundation via the Sinergia grant CRSII5−180359.


Translated image w/o the text module
Predicted depth w/o our text module
Translated image with our text module
Predicted depth with our text module
WHAT IN THE —/ IS THAT STRANGE LOOKING PLANE CHASING ME ?
Ground-truth Real Depth
Depth Estimator w/ Laplacian
Predicted real depth
Predicted Real Depth w/o text
Real w/o text
Generated real w/o text
A
Text-Adder Generator
B
M
FGAN
Depth Estimator w/ Laplacian
Predicted Comics Depth w/o text
$G_{R \to C}$
DUNIT
$G_{C \to R}$
Comics w/ text
L_{adv}
D_{text}
L_1
Predicted comics depth w/ text
A
B
Text-Adder = (1-M)*A + M*B
M
M→ $1 - M$
Generated comics w/o text
Comics w/o text
U-Net

Motivation for our text module

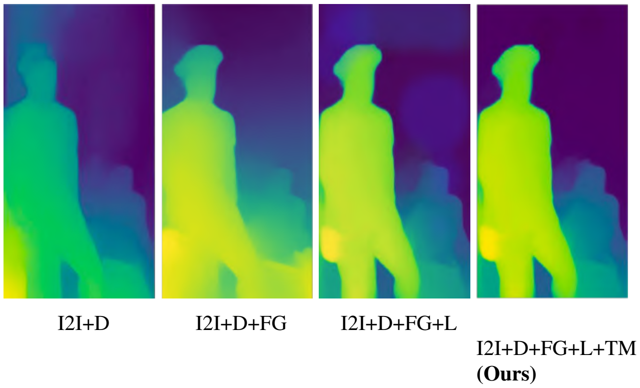
Overview of our approach with the text module

Figure 1: **Our depth estimation approach with the text-detection module.** We show, (Top): the motivation for our text module and (Bottom): the overall architecture of our approach incorporating the text module. In our text module, the masks are generated using a U-Net [9] trained on the text annotations form eBDtheque [3] dataset. The generated masks by the trained U-Net is used to train the text adder generator and the text adder 'ground-truth' as discussed above. The generated real image with text is then fed into the depth estimator. This predicts the depth with text. To remove the text based artifacts in the depth prediction, the complement of the text mask is multiplied with the predicted depth with text to, finally, predict the depth without the text.


IT'S THE LEADERS OF THE HELL GANG!
AND THE BLACK COBRA'S DONE IT AGAIN!
IT'S THE LEADERS OF THE HELL GANG!
AND THE BLACK COBRA'S DONE IT AGAIN!

Comics input MIDAS [8] CDE [5] Our

Figure 2: **Qualitative comparison of depth estimation** on the DCM validation images [7] from our benchmark. (Top Row): Depth predictions on the translated comics images as seen in the main paper. (Bottom Row): Depth predictions on the actual comics images (not translated). We show, from left to right, the input image in the comics domain, the result using the MIDAS [8] model, the result using the CDE [5] model, and **Our** model (comprising I2I, depth, feature GAN, Laplacian and the text module), respectively. We show that all the methods benefit from the I2I module. Further, we show that our approach can predict depth when applied both to the translated image, as well as the original comics image; while outperforming the baselines in both the scenarios. Cooler colors are farther and warmer colors are nearer (Best viewed in color).



I2I+D I2I+D+FG I2I+D+FG+L I2I+D+FG+L+TM **(Ours)**

Figure 3: **Qualitative comparison for the ablation study showing the effect of the different network components.** We show the depth predictions on the translated DCM validation images [7] from our benchmark. We report, from left to right, the depth predictions obtained by the model comprising I2I (DUNIT [2]) and Depth (CDE [5]) trained in an end-to-end manner; the result using the model comprising I2I, CDE and feature GAN; the result using the model comprising I2I, CDE, feature GAN and Laplacian; and **Our** model (comprising I2I, CDE, feature GAN, Laplacian and the text module), respectively. Cooler colors are farther and warmer colors are nearer (Best viewed in color).

References

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

[2] 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.

[3] 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.

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

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.