https://doi.org/10.2352/EI.2023.35.17.3DIA-104

Layered View Synthesis for General Images

Loïc Dehan, Wiebe Van Ranst and Patrick Vandewalle; EAVISE - PSI - ESAT, KU Leuven; Sint-Katelijne-Waver, Belgium

Abstract

We describe a novel method for monocular view synthesis. The goal of our work is to create a visually pleasing set of horizontally spaced views based on a single input image. This can be applied in view synthesis for virtual reality and glasses-free 3D displays.

Existing methods are able to form realistic results on images that show a clear distinction between a foreground object and the background. We aim to create novel views in more general scenarios, including more complex scenes in which there is no clear distinction. Our main contributions are in a computationally efficient method for realistic occlusion inpainting and blending, especially in complex scenes.

Our method can be effectively applied to any image, which is shown both qualitatively and quantitatively on a large dataset. Our method performs natural disocclusion inpainting and maintains the shape and edge quality of foreground objects.

Introduction

Augmented and virtual reality (AR/VR) head sets have become very popular in recent years. They offer the user a 3D perception of the scene by providing slightly different perspectives to the left and the right eye. This is actually using the same principle as we humans do for real-world scenes observed with our two eyes. Similarly, glasses-free 3D displays steer a separate view to each eye, allocating a subset of the display pixels to each view. Such displays are currently available commercially from companies like Leia, Looking Glass, Dimenco or Acer Spatial Labs.

Both for AR/VR head sets and glasses-free 3D displays, more than two views are actually often generated. This allows the viewer to move their head and (somewhat) look around an object. This effect, called motion parallax, adds to a natural 3D perception of the scene.

Meanwhile, when taking a picture, we want to be minimally bothered by additional requirements to capture the scene in 3D. Ideally, we would want to take a single 2D picture and then be able to visualize it in 3D.

In this paper, we present a method to create a multi-view image from a single 2D still picture. An example result is shown in Figure 1. Our focus is not on an accurate metric 3D reconstruction of the scene, but rather on a naturally-looking multiple-view set of images. As our target application is 3D visualization for a human observer, the perceptual image quality is more important than typical error metrics such as PSNR or MSE. While real-time processing is not required for our application with still pictures, it is important that our method can run efficiently on limited hardware (e.g. a tablet or smartphone).

Our work is based on the SLIDE algorithm by Jampani et al. [8]. We aim to generate multiple horizontally spaced views from a single 2D picture without visually disturbing artifacts. As our main application is visualization on glasses-free 3D displays, we focus here on horizontally spaced views (although most of the work could be applied to vertically spaced views as well). Our main contributions over the current state-of-the-art are maintaining edge quality, a more natural inpainting of disocclusions around foreground objects and a robust method to handle disocclusions at multiple depth levels.

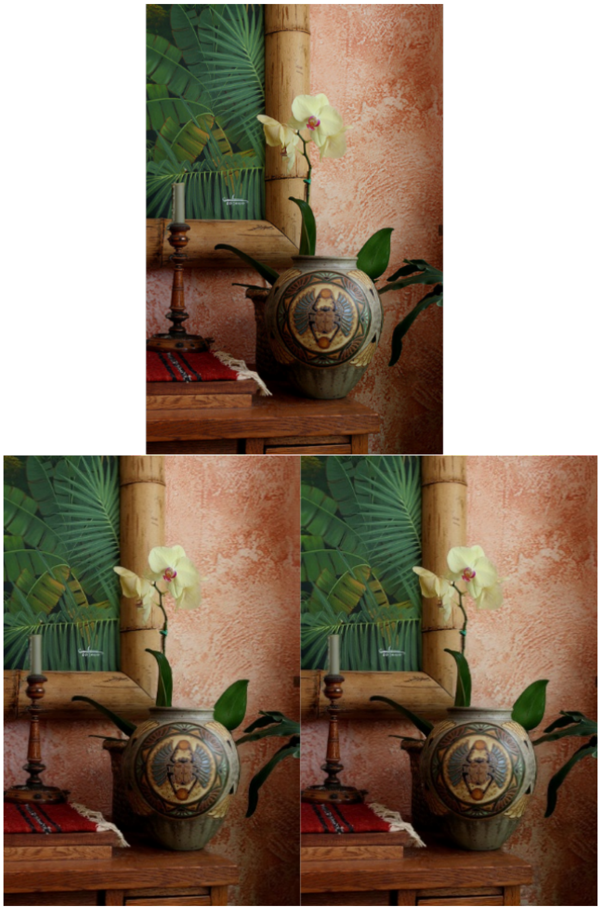

Figure 1. *Image from Holopix50k [5] dataset with two horizontally spaced novel views generated using our proposed algorithm.*

The rest of this paper is structured as follows: First we discuss related work. Next, we describe the dataset used in our experiments. Our algorithm is then presented in the following section, and results will be shown. Finally, we draw some conclusions.

Related Work

Recently, multiple approaches have been proposed for view synthesis from a 2D image. An algorithm 3D-Photo for rendering novel views from a single RGB-D image was presented by Shih et al. [17]. They convert the RGB-D image into a layered depth image (LDI) and apply inpainting of the background around foreground objects to fill in occlusion areas. Jampani et al. presented SLIDE, adding a soft layering approach to create more natural effects around hair and allow matting effects [8].

Most approaches for novel view synthesis can be subdivided in the following steps: first, a depth map is estimated for the 2D image using monocular depth estimation; next, the 3D scene is completed using occlusion inpainting and similar techniques; finally, novel views can be rendered from the completed image using view synthesis. Optionally, some post-processing is performed to enhance the generated views. Related work on each of these topics will be further discussed in the following subsections.

Mildenhall et al. presented an alternative solution in their breakthrough paper on neural radiance fields (NeRF) [11]. Using NeRF, a neural network is trained to generate novel views of a scene from a set of input views of the scene. Very impressive results have been obtained using this method, with a wide variety of follow-up works. Typically, this approach is used to generate novel views with a wide baseline (wider than we aim here). Moreover, neural radiance fields require a large number of input images, and a neural network to be trained per scene, which is computationally much more demanding than our objective here.

In the next subsections, we will describe some relevant approaches to depth estimation and image completion.

Depth Estimation

Monocular depth estimation techniques aim to estimate dense depth [1, 18, 24] based on a single RGB image. Whereas a wide range of algorithms was developed using various heuristics (focus, edge analysis, typical organisation of scenes, etc.), most recent methods apply deep learning to train depth estimation from large datasets. Many methods directly utilize a single image [3, 9, 16] or estimate an intermediate 3D representation such as point clouds [19, 21]. Some other methods combine an RGB image with, for example, sparse depth maps or normal maps [10, 15] to estimate dense depth maps. These methods are trained on large scale datasets generated from RGB-D cameras, thus they can only reproduce the raw depth scan. For our purpose, we need to generate a dense depth map for a general image in a computationally efficient way. The Midas algorithm proposed by Ranftl et al. [16] is currently most suited for our goals. They used a multiscale ResNeXt architecture that was pretrained on ImageNet. A new loss function was applied for training, using a scale- and shift-invariant term based on absolute depth error and a regularization term to push depth discontinuities to be sharp and coincide with object edges. The main contribution of this work was their use of a broad collection of datasets for training and testing, rather than a single dataset. Ranftl et al. trained their algorithm on a collection of widely varying datasets and supplemented those with another large dataset based on frames from (stereoscopic) 3D movies.

Image Completion

Recent deep learning techniques can realistically complete a set of masked regions in an image. These methods are based on the advent of Generative Adversarial Networks (GAN) [4]. Pathak et al. [13] introduce an adversarial loss in addition to the reconstruction loss to address that inpainting is multimodal. Iizuka et al. [7] formed improvements by introducing both global and local discriminators for deriving the adversarial losses. More recently, Yu et al. [22] presented a contextual attention mechanism in a generative inpainting framework, which further improves the inpainting quality. Nazeri et al. [12] observed that the structure of an image is represented in the edge map. They achieve photorealistic results by first completing the edges before completing the actual pixels. Zhao et al. [25] use aspects from style transfer research [2, 6] to introduce co-modulated GANs. Their method can generate realistic results on larger mask sizes.

Dataset

Our method is focused on creating horizontally spaced views. Therefore, we evaluate our method on stereo datasets. The Holopix50k [5] dataset comprises 49,368 image pairs contributed by users of the Holopix™ mobile social platform. It covers a wide range of scenes. This large variety of diverse scenarios is very representative for the variation in digital still pictures, and can significantly improve the generalization of deep models.

Our approach

Our method broadly follows the SLIDE method [8]. It can be subdivided in the following steps:

1. Based on a single RGB image, we first perform depth estimation using the Midas algorithm [16].
2. Secondly, we form a full-resolution inpainting mask that highlights the areas we need to inpaint in order to later fill disocclusions. Our inpainting mask takes multi-level disocclusions into account. We dilate the depth map and determine blending values to prevent striping and dilation artifacts.
3. Next, we perform inpainting. We modified the LaMa inpainting method from Zhao et al. [25] to inpaint RGB-D images and split the training inpainting masks into disocclusion masks and large random masks to improve inpainting quality.
4. Finally, we render both the regular foreground RGB-D image and the inpainted background RGB-D image using a forward warping method. The output image is obtained by filling the disocclusion holes in the foreground image using the rendered background image.

The following subsections will describe our method in more detail.

Depth estimation

Initially, we estimate a depth map of the input image using the Midas monocular depth estimation method by Ranftl et al. [16]. Their method is applicable to a wide range of images at a high resolution. However, near the edges of foreground objects, their method cannot form a crisp, stepwise transition from the background to the foreground depth. Some transitional depth values are visible near the edges of objects, as shown in Figure 2.

These transitional depth values near the edges of objects cause visual artifacts when rendering novel views using forward or backward mapping. Jampani et al. [8] describe the formation of stretched triangles near the edges of foreground objects after backward mapping. When using forward mapping, a striping artifact is created. The transitional depth values are each mapped to

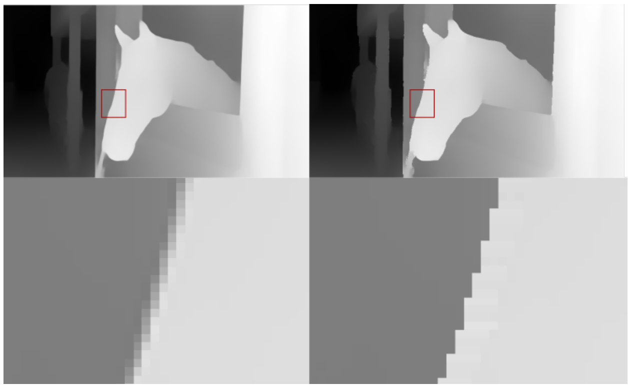

Figure 2. Top left: depth map generated by Midas. Top right: dilated depth map. Bottom left and right: top images zoomed in on the area indicated by the red rectangle.

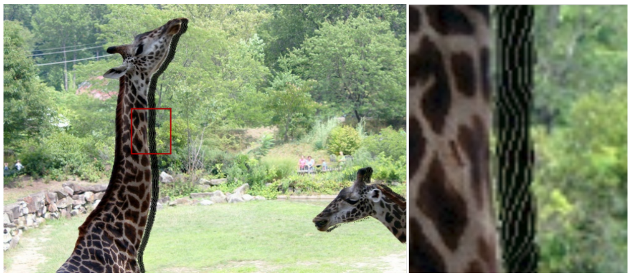

Figure 3. Example rendering without disocclusion inpainting to highlight striping. Left: full image. Right: zoomed in part of the image indicated by the red rectangle.

a slightly further position, spread across the disoccluded area. Additionally, the edge of the foreground object is damaged as some of the pixels near the edge have been mapped away from the rest of the object. An example of this striping artifact is illustrated in Figure 3.

SLIDE [8] avoids these artifacts by using a combination of the image segmentation network by Qin et al. [14] and their soft foreground visibility. However, this approach only works in a scene where a single subject is portrayed and if there is a clear two-layer distinction between the background and the foreground object. In most images, the scene is more cluttered, and it is impossible to form a clear (binary) separation into a foreground and background layer. Secondly, when depth estimation and image segmentation are separated into two independent networks, additional artifacts may occur when they are drastically different, as depicted in Figure 4.

We propose to avoid the striping artifacts by dilating the depth map. Firstly, we determine the local minimum and maximum of the transitional depth values. We set the value of each of the transitional depth values to the maximum. Secondly, we determine a blending value based on the original, minimum and maximum depth values. Finally, the blending value is used to transition between the rendered foreground and background layer of the image. Through the blending operation, we remove the dilation artifacts visible after rendering.

Our approach does not rely on an explicit separation between a foreground and background as it can be applied to any depth transition in the depth map. To limit the computational demand of this method, it is only applied to pixels where the gradient of the depth map indicates a significant transition.

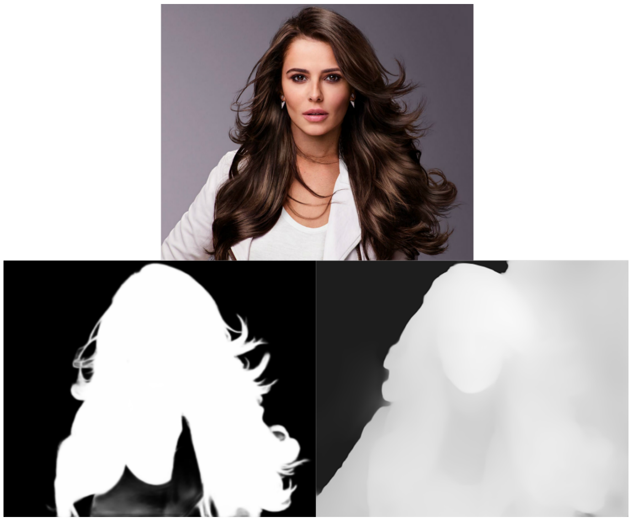

Figure 4. Image from Qin et al. [14] with the corresponding foreground estimation in the bottom left and depth estimation in the bottom right.

Inpainting mask

We need to inpaint areas that will potentially be disoccluded when the viewpoint is changed. SLIDE [8] determines an inpainting mask by calculating for each pixel whether pixels in the local neighborhood change more in depth than position. Since this is highly computationally expensive, they initially downscale the image. After upscaling the generated mask, the resulting inpainting mask is far broader than the required inpainted region and has a blocked-out edge. This reduces the quality of the inpainted background layer. This approach also assumes a clear, singular transition from a background layer to a foreground object. As mentioned previously, most images have a more cluttered scene in which there are often multiple transitions close to each other. This two-layer rendering approach, therefore, causes multi-layer disocclusion artifacts.

For our purpose, we are generating horizontally spaced views, and we can therefore focus on horizontal depth transitions in the dilated depth map. We iterate over the dilated depth map, and whenever a sudden increase/decrease is reached, the pixels on the higher side of this transition are masked. An example of the different masks is shown in Figure 5.

This allows us to keep track of a reference depth value to avoid multi-level disocclusion artifacts. This reference depth value can be used to verify if the inpainted area corresponds with the expected depth in that region. If the inpainted background depth is further than the expected reference depth, there is likely a multi-level-disocclusion artifact. When this is the case, that part of the background layer is not used for the disocclusion inpainting. Instead, we use simple reflection inpainting to fill these remaining holes.

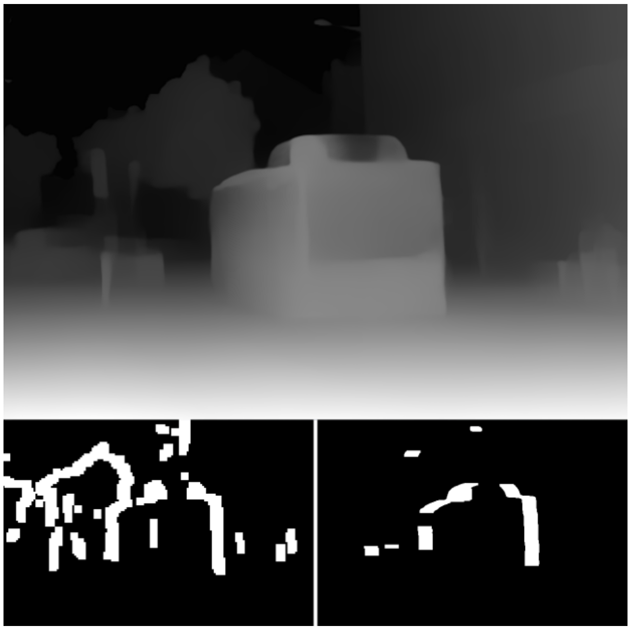

Figure 5. Top: depth map generated using Midas [16]. Bottom-left: mask generated as described in SLIDE by Jampani et al. [8]. Bottom-right: our inpainting mask.

Inpainting

The inpainting mask, determined in the previous section, indicates areas that will potentially be disoccluded when the viewpoint is changed. To fill these disocclusion holes, we first form a background image by inpainting these areas. Jampani et al. [8] indicate in their SLIDE paper that training the inpainting network on disocclusion masks instead of only randomly generated masks improves the inpainting quality for this case. However, more crowded scenes and multi-level disocclusions can cause large masked regions to appear. We train the network on a combination of random inpainting masks and disocclusion masks as opposed to only disocclusion masks. Most of the general inpainting can be learned from random inpainting masks. In this way, the network can also handle larger masks that may occur around multi-level disocclusions. Additionally, we replace DeepFillv2 with the LaMa inpainting network by Zhao et al. [25] as this is more suited for high resolution and large-mask inpainting. Similarly to SLIDE [8] we modify the network for RGB-D inpainting.

Rendering

We have chosen to implement our rendering method using forward mapping instead of backward mapping as used in SLIDE. Initially, the foreground and background RGB-D images are both rendered separately for the novel viewpoint. Secondly, we fill the disocclusion holes in the foreground rendering with the information from the background rendering in the same position. Next, the blending values determined in the depth estimation are used to remove the dilation artifacts and create a smooth transition between the foreground and background rendering.

Finally, reflection inpainting is used to fill the holes left near the borders of the image, since neither the foreground nor the background layer will be mapped to these areas. Since these remaining holes are relatively small and near the edge of the image, we choose to use simple reflection inpainting in these remaining areas.

Experimental Results

In the previous sections, we described our method for novel view synthesis from a single 2D input image. We compare our method to the SLIDE view synthesis method by Jampani et al. [8], to the SynSin method by Wiles et al. [20] and to the 3D-Photo method by Shih et al. [17] using the following four evaluation metrics:

1. Mean squared error (MSE);
2. Peak signal-to-noise ratio (PSNR);
3. Structural similarity index measure (SSIM);
4. Learned perceptual image patch similarity (LPIPS) [23].

We apply these metrics to the Holopix50k [5] dataset. Figure 6 shows a result for the different methods. The results are summarized in Table 1.

Method	MSE↓	PSNR ↑	SSIM ↑	LPIPS [23] ↓
SynSin	1639.46	16.88	0.458	0.374
3D-Photo	955.43	19.42	0.566	0.152
SLIDE	949.54	19.46	0.568	0.121
Ours	**938.50**	**19.59**	**0.571**	**0.111**

Table 1. Evaluation of our method compared to the SLIDE view synthesis method by Jampani et al. [8], the SynSin method by Wiles et al. [20] and to the 3D-Photo method by Shih et al. [17] on the Holopix50k [5] dataset.

We notice SynSin [20] has the worst evaluation. This is because this network is limited to images of lower resolution (256 × 256). Secondly, the internal depth estimation formed by their method is not as accurate as Midas causing objects in the final rendering not to be aligned with the ground truth.

The 3D-Photo method by Shih et al. [17] can form realistic disoclusion completions but also causes severe deformations as highlighted in Figure 6. It is also worth noting that this method is significantly slower than the other three.

In comparison to the state-of-the-art SLIDE method [8], our method reduces more striping/stretching artifacts, especially in crowded or complex scenes. Our improvements to the inpainting mask and network also lead to more realistic completions surrounding foreground objects. An example is illustrated in Figure 7.

Even more importantly than the quantitative comparisons in Table 1, we also evaluated our approach qualitatively on a Leia LumePad glasses-free multi-view 3D tablet and confirmed that our approach produces natural and visually pleasing results.

Conclusion

We have presented a novel method for view synthesis creating visually pleasing multi-view images. Our method creates horizontally spaced images. We improve upon the state-of-the-art method by reducing striping, stretching and multi-level disocclusion artifacts. We maintain edge quality and form smooth disocclusion inpainting results using a matting approach. Both quantitative and qualitative evaluations show that our method outperforms state-of-the-art methods, especially for complex images.

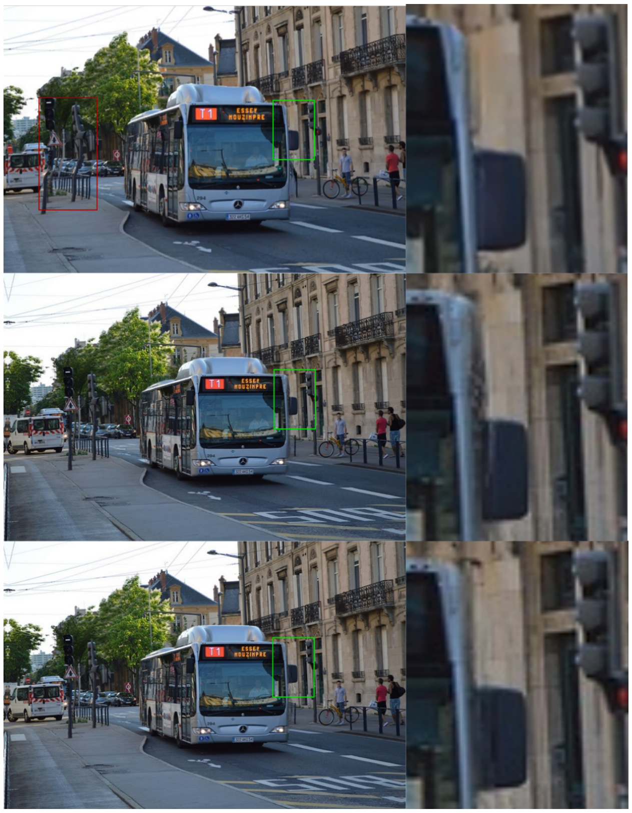

Figure 6. Visual comparison of a rendering using the following methods: (top) 3D-Photo [17], (middle) SLIDE [8], (bottom) ours. Red rectangle in the top image highlights deformation. Green rectangle indicates the area magnified on the right.

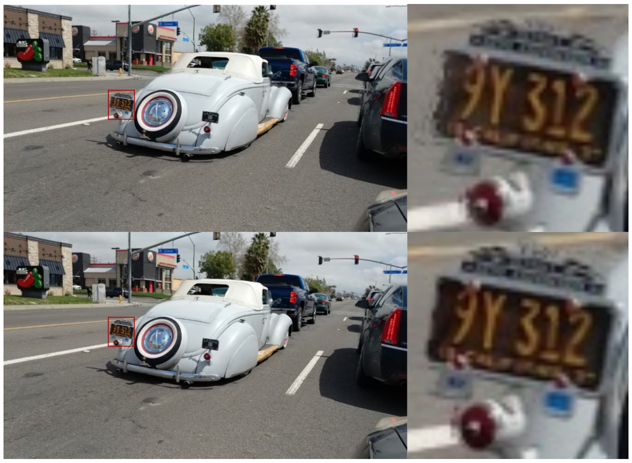

Figure 7. Visual comparison of a rendering using the following methods: (top-left) SLIDE [8], (bottom-left) ours. Top- and bottom right show the zoomed in part of the images indicated by the red rectangle.

Acknowledgement

We would like to thank Leia Inc for funding this work.

References

[1] Amlaan Bhoi. Monocular depth estimation: A survey. In *arXiv preprint arXiv:1901.09402*, 2019.

[2] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *ICLR*, 2017.

[3] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proc. IEEE international conference on computer vision*, pages 3828–3838, 2019.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, oct 2020.

[5] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, Seattle, WA, June 2020.

[6] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[7] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):107:1–107:14, 2017.

[8] Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T Freeman, David Salesin, Brian Curless, et al. SLIDE: Single image 3D photography with soft layering and depth-aware inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12518–12527, 2021.

[9] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4521–4530, 2019.

[10] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *Proc. International Conference on Robotics and Automation (ICRA)*, pages 3288–3295, 2019.

[11] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020.

[12] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. EdgeConnect: Structure guided image inpainting using edge prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[13] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 06 2016.

[14] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020.

[15] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In . *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2019.

[16] 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*, 44(3), 2022.

[17] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020.

[18] Lokender Tiwari, Pan Ji, Quoc-Huy Tran, Bingbing Zhuang, Saket Anand, and Manmohan Chandraker. Pseudo rgb-d for self-improving monocular slam and depth prediction. In *Proc. European Conference on Computer Vision*, pages 437–455, 2020.

[19] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. In *Proc. IEEE International Conference on Computer Vision Workshops*, 2019.

[20] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020.

[21] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In *arXiv preprint arXiv:1906.06310*, 2019.

[22] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4470–4479, 2019.

[23] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[24] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. In *Science China Technological Sciences*, 2020.

[25] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I-Chiao Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2021.

Author Biography

Loïc Dehan received his MSc in engineering technology from KU Leuven (2021). He worked as a researcher at KU Leuven until 2022.

Wiebe Van Ranst received his MSc in engineering technology from KU Leuven (2013). He did a PhD on real-world applications of AI on constrained hardware at KU Leuven (2019). He is now a post-doctoral researcher at KU Leuven. His work focuses on embedded deep learning algorithms for computer vision applications.

Patrick Vandewalle received a MSc degree in electrical engineering from KU Leuven (2001), and a PhD degree from EPFL on super-resolution imaging (2006). From 2007 to 2018, he worked at Philips Research as a senior research scientist. He is now an associate professor at KU Leuven. His current research in the EAVISE research group focuses on 3D processing, reconstruction, computer vision and AR/VR.