Machine Vision and Applications (2023) 34:82
https://doi.org/10.1007/s00138-023-01410-5

ORIGINAL PAPER

Unsupervised single-shot depth estimation using perceptual reconstruction

Christoph Angermann[1] · Matthias Schwab[1] · Markus Haltmeier[1] · Christian Laubichler[2] · Steinbjörn Jónsson[3]

Received: 3 August 2022 / Revised: 10 March 2023 / Accepted: 23 May 2023

Abstract
Real-time estimation of actual object depth is an essential module for various autonomous system tasks such as 3D reconstruction, scene understanding and condition assessment. During the last decade of machine learning, extensive deployment of deep learning methods to computer vision tasks has yielded approaches that succeed in achieving realistic depth synthesis out of a simple RGB modality. Most of these models are based on paired RGB-depth data and/or the availability of video sequences and stereo images. However, the lack of RGB-depth pairs, video sequences, or stereo images makes depth estimation a challenging task that needs to be explored in more detail. This study builds on recent advances in the field of generative neural networks in order to establish fully unsupervised single-shot depth estimation. Two generators for RGB-to-depth and depth-to-RGB transfer are implemented and simultaneously optimized using the Wasserstein-1 distance, a novel perceptual reconstruction term, and hand-crafted image filters. We comprehensively evaluate the models using a custom-generated industrial surface depth data set as well as the Texas 3D Face Recognition Database, the CelebAMask-HQ database of human portraits and the SURREAL dataset that records body depth. For each evaluation dataset, the proposed method shows a significant increase in depth accuracy compared to state-of-the-art single-image transfer methods.

Keywords Unsupervised learning · Wasserstein GAN · Surface depth · Perceptual similarity

1 Introduction

Real-time depth inference of a given object is an essential computer vision task which can be applied in various robotic tasks such as simultaneous localization and mapping [1–3] as well as autonomous quality inspection in industrial applications [4, 5]. As the popularity of VR applications has continued to grow, instant depth estimation has also become an integral part of modeling complex 3D information out of single 2D images of human faces [6, 7] or body parts [8–10]. Depth information about an object can be directly obtained from sensors for optical distance measurement. Time-of-Flight (ToF) cameras, LIDAR or stereo imaging systems are often used in practice and were also employed to generate paired RGB-depth data from some well-known depth databases [1, 2, 8, 10–13]. Since these sensors are typically costly and time-consuming devices that are also sensitive to external influences, their applicability to fast full-image depth generation on small on-site devices is limited. These limitations have motivated depth synthesis out of a simpler modality in terms of acquisition effort, namely a conventional RGB image. This development has initiated a completely new field of research in computer vision.

An important contribution in that area was made by Eigen et al. [14], who proposed deep convolutional neural networks (DCNNs) for monocular depth synthesis of indoor and outdoor scenes. Basically, monocular single-image depth estimation out of RGB images can be seen as a modality

✉ Christoph Angermann
christoph.angermann@student.uibk.ac.at

Matthias Schwab
matthias.schwab@uibk.ac.at

Markus Haltmeier
markus.haltmeier@uibk.ac.at

Christian Laubichler
christian.laubichler@lec.tugraz.at

Steinbjörn Jónsson
steinbjorn.jonsson@innio.com

[1] Department of Mathematics, University of Innsbruck, Technikerstraße 13, 6020 Innsbruck, Austria

[2] LEC GmbH, Inffeldgasse 19, 8010 Graz, Austria

[3] INNIO Jenbacher GmbH & Co OG, Achenseestrasse 1-3, 6200 Jenbach, Austria

Published online: 11 August 2023

transfer in which observed data of one modality is mapped to desired properties of another, potentially more complex, modality. Although DCNNs are promising tools that succeed on such transfer tasks, they are commonly based on large amounts of training data, and generation and acquisition can be a demanding task. In the supervised setting in particular, DCNNs make use of paired training data during network parameter optimization, i.e., the network is provided with a single-view RGB and corresponding per-pixel depth [6, 10, 14, 15]. Since large-scale dense depth profiles are not abundant in many applications, supervised approaches are not feasible for these objects. One possible way to address these shortcomings of supervised methods is to consider self-supervised approaches based on monocular video clips in which a supervisory depth counterpart is extracted from pose changes between adjacent frames.

These models can be trained on RGB sequences in a self-supervised manner, where a depth network and a pose estimation network are simultaneously optimized via sophisticated view-synthesis losses [3, 16–18]. Obviously, these methods require non-static scenes or a moving camera position (e.g., moving humans [18], autonomous driving [2]).

A very recent example for a scenario, where neither video sequences, stereo pairs nor paired data are available, is non-destructive evaluation of internal combustion engines for stationary power generation [4, 5]. Within this application, surface depth information has to be extracted from RGB image data. With current standards, cylinder condition can be assessed from a depth profile on a micrometer scale of the measured area (cf. Fig. 1). However, microscopic depth sensing of cylinder liner surface areas is a time-consuming and resource-intensive task which consists of disassembling the liner, removing it from the engine, cutting it into segments and measuring them with a highly expensive and stationary confocal microscope [4]. With a handheld microscope, however, single RGB records of the liner's inner surface can be generated from which depth profiles may be synthesized. Since depth data is generated on a quite small scale (1.9 mm $\times$ 1.9 mm) and is comparatively high resolved, it is hardly possible to generate RGB data with accurately aligned pixel positions. This results in an unsupervised approach required for reasonable depth synthesis of this static scene.

The main objective of this study is to propose a general method for depth estimation out of scenes for which neither paired data, video sequences, nor stereo pairs are available. For this, we consider the depth estimation problem as an intermodal transfer task of single images. Several recent advances in unpaired modality transfer are based on generative adversarial models (GAN) [19], cycle-consistency [20] and probabilistic distance measures [21, 22]. The method proposed in this paper builds on established model architectures and training strategies in deep learning which are beneficially combined for unpaired single-view depth synthesis. Introduction of a novel perceptual reconstruction term in combination with appropriate hand-crafted filters further improves accuracy and depth contours.

The method is comprehensively tested on the afore mentioned industrial application of surface depth estimation. Additionally, the approach is applied to other, external, datasets to create realistic scenarios where perfectly aligned RGB-depth data of single images is not available in practice. More precisely, we test the model on the Texas 3D Face Recognition database (Texas-3DFRD) [12], the Bosphorus-3DFA [11] and the CelebAMask-HQ [23] to show its plausibility for facial data in an unsupervised setting.

The SURREAL dataset [9] is used to test performance on RGB-D videos of human bodies, where RGB and depth frames are not perfectly aligned. For every evaluation experiment, the depth accuracy of the proposed framework is compared to state-of-the-art methods in unsupervised single-image transfer. To be more precise, the methods used for comparison are standard cycleGAN [20], CUT [24] that uses contrastive learning for one-sided transfer and gcGAN [25] that utilizes geometric constraints between modalities. For facial data, we additionally compare to Wu et al. [26], a very recent work where in addition to the depth profile also the albedo image, the illumination source and a symmetry confidence map is predicted in an unsupervised manner.

Contributions:

- This study finds a solution to the industrial problem of single-shot surface depth estimation where no paired data, no video sequences and no stereo pairs are available.
- In this work, depth estimation is considered as a single-image modality transfer; the proposed method shows superior performance over state-of-the-art works, quantitatively and qualitatively.
- Application to the completely different tasks of unsupervised face and human body depth synthesis indicates the universality of the approach.

2 Related work

The following section summarizes important milestones in the development of generative adversarial networks and highlights important work on single-image depth estimation as well as depth synthesis via GANs. In the supplementary, background is provided on some 3D databases that have been critical to the development of deep learning-based models for depth estimation.

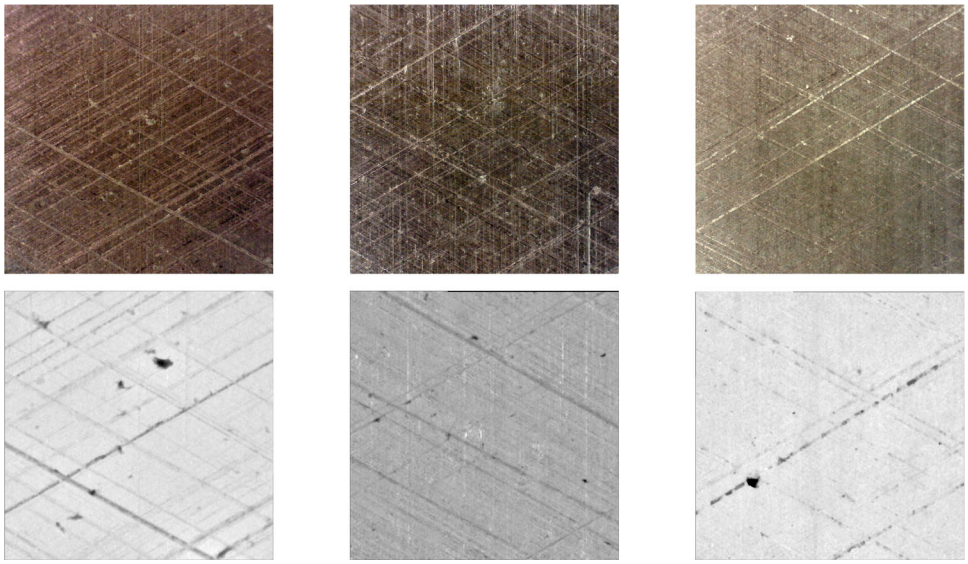

Fig. 1 Top: RGB measurements of the inner surface of three cylinder liners with a spatial range of 4.2 mm × 4.2 mm, recorded by a handheld microscope. Bottom: Depth profile of the same cylinder with a spatial range of 1.9 mm × 1.9 mm, measured with a confocal microscope. The pixels of the modalities are not aligned

2.1 Generative adversarial networks

A standard GAN [19] consists of a generator network $G\colon \mathcal{Z} \to \mathcal{X}$ mapping from a low-dimensional latent space $\mathcal{Z}$ to image space $\mathcal{X}$, where parameters of the generator are adapted so that the distribution of generated examples assimilates the distribution of a given data set. To be able to assess any similarity between arbitrary high-dimensional image distributions, a discriminator $f\colon \mathcal{X} \to [0, 1]$ is trained simultaneously to distinguish between generator distribution and real data distribution. In a two-player min-max game, generator parameters are then updated to fool a steadily improving discriminator.

Usage of the initially proposed discriminator approach can cause the vanishing gradient problem and does not provide any information on the real distance between the generator and the real distribution. This issue has been discussed thoroughly in [21], where the problem is bypassed by replacing the discriminator with a critic network that approximates the Wasserstein-1 distance [27] between the real distribution and the generator distribution.

While the quintessence of GANs is to draw synthetic instances following a given data distribution, cycle-consistent GANs [20] allow one-to-one mappings between two image domains $\mathcal{X}$ and $\mathcal{Y}$. In essence, two generator networks $G_{\mathcal{Y}}\colon \mathcal{X} \to \mathcal{Y}$, $G_{\mathcal{X}}\colon \mathcal{Y} \to \mathcal{X}$ and corresponding discriminator networks $f_{\mathcal{Y}}\colon \mathcal{Y} \to [0, 1]$, $f_{\mathcal{X}}\colon \mathcal{X} \to [0, 1]$ are trained simultaneously to enable generation of synthetic instances for both image domains (e.g., synthesizing winter landscapes from summer scenes and vice versa). To ensure one-to-one correspondence, a cycle-consistency term is added to the two adversarial loss functionals. While cycle-consistent GANs had initially been constructed for style transfer purposes, they were also very well received in the area of modality transfer in biomedical applications [28–30]. Since optimization and fine-tuning of GANs often turns out to be extremely demanding and time-intensive, much research has emphasized stabilization of the training process through the development of stable network architectures such as DCGAN [31] or PatchGAN [32].

2.2 Monocular depth estimation

Deep learning-based methods achieve state-of-the-art results on depth synthesis task by training a DCNN on a large-scale and extensive data set [1, 2]. Most of RGB-based models are supervised, i.e. they require corresponding depth data that is pixel-wise aligned. One of the first DCNN approaches by Eigen et al. [14] included sequential deployment of a coarse-scale stack and a refinement module and was benchmarked on the KITTI [2] and the NYU Depth v2 data set [1]. Using an encoder–decoder structure in combination with an adversarial loss term helped to increase visual quality of the dense depth estimates [33]. Later methods also considered deep residual networks [34] or deep ordinal regression networks [35] in order to significantly increase performance on these data sets, where commonly considered performance measures are the root mean squared error (RMSE) or the δ_1 accuracy [3]. Since a lot of research focused on further performance increase at the expense of model complexity and runtime, Wofk et al. [36] used a lightweight network architecture [37] and achieved comparable results.

2.3 Depth estimation using GAN

Use of left-right consistency and a GAN architecture results in excellent unsupervised depth estimation based on stereo images [38, 39]. In [40] and [41], a GAN has been trained

to perform unpaired depth synthesis out of single monocular images. To this end, GANs were employed in the context of domain adaptation using an additional synthesized data set of the same application with paired samples. This approach may not be regarded as a fully unsupervised method and requires availability or construction of a synthetic dataset. Arslan et Seke [6] consider a conditional GAN (CGAN) [32] for solving single-image face depth synthesis. Nevertheless, CGANs rely on paired data since the adversarial part estimates the plausibility of an input–output pair. Another interesting approach was taken in [15], where indoor depth and segmentation were estimated simultaneously using cycle-consistent GANs. The cycle-consistency loss helped them to maintain the characteristics of the RGB input during depth synthesis, while the simultaneous segmentation resolved the fading problem in which depth information is hidden by larger features. However, the proposed discriminator network and reconstruction term in the generator loss function are based on paired RGB and depth/segmentation data, which is not available for the aforementioned industrial application of surface depth synthesis.

3 Method

This section proposes an approach to monocular single-image depth synthesis with unpaired data and discusses the introduced framework and training strategy in detail.

3.1 Setting and GAN architecture

The underlying structure of the proposed modality synthesis is two GANs linked with a reconstruction term (cf. Fig. 2). To be more exact, let $\mathcal{X} \subset [0, 255]^{d_1 \times d_2 \times 3}$ and $\mathcal{Y} \subset \mathbb{R}^{d_1 \times d_2 \times 1}$ denote the domain of RGB and depth images, respectively, where the number of image pixels $d_1 \cdot d_2$ is the same in both domains. Furthermore, let $X{:=}\{x_1, \ldots, x_M\}$ be the set of M given RGB images and $Y{:=}\{y_1, \ldots, y_N\}$ the set of N available but unaligned depth profiles. $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ denote the distributions of the images in both domains. The proposed model includes a generator function $G_{\theta_{\mathcal{Y}}} \colon \mathcal{X} \to \mathcal{Y}$, which aims to map an input RGB image to a corresponding depth counterpart in the target domain. A generator function for image transfer may be approximated by a DCNN, which is parameterized by a weight vector $\theta_{\mathcal{Y}}$ consisting of several convolution kernels. By adjusting $\theta_{\mathcal{Y}}$, the distribution of generator outputs $P_{\theta_{\mathcal{Y}}}$ may be brought closer to the real data distribution in the depth domain $P_{\mathcal{Y}}$. Note we do not know what $P_{\theta_{\mathcal{Y}}}$ and $P_{\mathcal{Y}}$ actually look like, we only have access to unpaired training samples $G_{\theta_{\mathcal{Y}}}(x) \sim P_{\theta_{\mathcal{Y}}},\ x \in X$ and $y \sim P_{\mathcal{Y}},\ y \in Y$. An adversarial approach is deployed to ensure assimilation of both high-dimensional distributions in the GAN setting. The distance between the generator distribution and the real distribution is estimated by an additional DCNN $f_{\omega_{\mathcal{Y}}} \colon \mathcal{Y} \to \mathbb{R}$, which is parameterized by weight vector $\omega_{\mathcal{Y}}$ and is trained simultaneously with the generator network since $P_{\theta_{\mathcal{Y}}}$ changes after each update to the generator weights $\theta_{\mathcal{Y}}$. This ensures that $G_{\theta_{\mathcal{Y}}}$ can be pitted against a steadily improving loss network $f_{\omega_{\mathcal{Y}}}$ [19].

This research work has chosen a network critic based on the Wasserstein-1 distance [21, 27]. The Wasserstein-1 distance (earth mover distance) between two distributions P_1 and P_2 is defined as $\mathcal{W}_1(P_1, P_2){:=}\inf_{J \in \mathcal{J}(P_1, P_2)} \mathbb{E}_{(x,y)\sim J} \|x - y\|$, where the infimum is taken over the set of all joint probability distributions that have marginal distributions P_1 and P_2. Since the exact computation of the infimum is highly intractable, the Kantorovich–Rubinstein duality [27] is used

$$\mathcal{W}_1(P_1, P_2) = \sup_{\|f\|_L \leq 1} \left[\mathop{\mathbb{E}}_{y\sim P_1} f(y) - \mathop{\mathbb{E}}_{y\sim P_2} f(y) \right], \tag{1}$$

where $\|\cdot\|_L \leq C$ denotes that a function is C-Lipschitz. Equation (1) indicates that a good approximation to $\mathcal{W}_1(P_{\mathcal{Y}}, P_{\theta_{\mathcal{Y}}})$ is found by maximizing the distance $\mathbb{E}_{y\sim P_{\mathcal{Y}}} f_{\omega_{\mathcal{Y}}}(y) - \mathbb{E}_{y\sim P_{\theta_{\mathcal{Y}}}} f_{\omega_{\mathcal{Y}}}(y)$ over the set of DCNN weights $\{\omega_{\mathcal{Y}} \mid f_{\omega_{\mathcal{Y}}} \colon \mathcal{Y} \to \mathbb{R}\ \text{1-Lipschitz}\}$, where the Lipschitz continuity of $f_{\omega_{\mathcal{Y}}}$ can be enhanced via a gradient penalty [22]. Given training batches $\mathbf{y} = \{y_n\}_{n=1}^b,\ y_n \overset{\text{iid}}{\sim} P_{\mathcal{Y}}$ and $\mathbf{x} = \{x_n\}_{n=1}^b,\ x_n \overset{\text{iid}}{\sim} P_{\mathcal{X}}$, this yields the following empirical risk for critic $f_{\omega_{\mathcal{Y}}}$:

$$\begin{aligned}\mathcal{R}_{\text{cri}}(\omega_{\mathcal{Y}}, \theta_{\mathcal{Y}}, p, \mathbf{y}, \mathbf{x}) := \frac{1}{b} \sum_{n=1}^{b} \Big[& f_{\omega_{\mathcal{Y}}}(G_{\theta_{\mathcal{Y}}}(x_n)) - f_{\omega_{\mathcal{Y}}}(y_n) \\ & + p \cdot \Big(\big(\|\nabla_{\tilde{y}_n} f_{\omega_{\mathcal{Y}}}(\tilde{y}_n)\|_2 - 1 \big)_+ \Big)^2 \Big],\end{aligned} \tag{2}$$

where p denotes the influence of the gradient penalty, $(\cdot)_+ := \max(\{0, \cdot\})$ and $\tilde{y}_n := \epsilon_n \cdot G_{\theta_{\mathcal{Y}}}(x_n) + (1 - \epsilon_n) \cdot y_n$ for $\epsilon_n \overset{\text{iid}}{\sim} \mathcal{U}[0, 1]$. The goal of the RGB-to-depth generator $G_{\theta_{\mathcal{Y}}}$ is to minimize the distance. Since only the first term of the functional in (2) depends on the generator weights $\theta_{\mathcal{Y}}$, the adversarial empirical risk for generator $G_{\theta_{\mathcal{Y}}}$ simplifies as follows:

$$\mathcal{R}_{\text{adv}}(\theta_{\mathcal{Y}}, \omega_{\mathcal{Y}}, \mathbf{x}) := -\frac{1}{b} \sum_{n=1}^{b} f_{\omega_{\mathcal{Y}}}(G_{\theta_{\mathcal{Y}}}(x_n)). \tag{3}$$

3.2 Perceptual reconstruction

In the context of depth synthesis, it is not sufficient to ensure that the output samples lie in the depth domain. Care must be taken that synthetic depth profiles do not become irrelevant to the input. A reconstruction constraint forces generator input and output to share same spatial

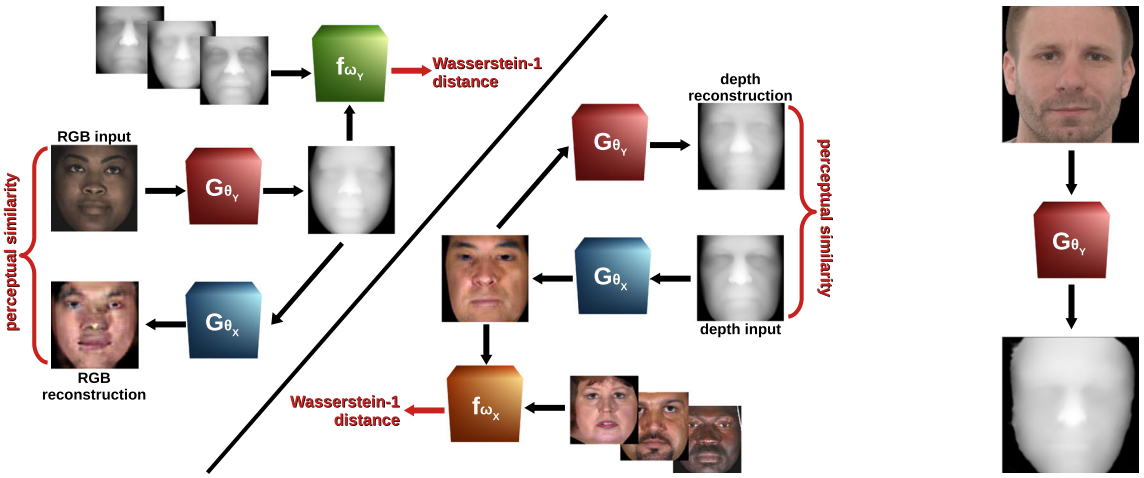

Fig. 2 Illustration of the proposed framework: The left part describes the domains in which the RGB-to-depth generator $G_{\theta_\mathcal{Y}}$ and the contrary depth-to-RGB generator $G_{\theta_\mathcal{X}}$ operate. Both generators are updated via the probabilistic Wasserstein-1 distance, estimated by $f_{\omega_\mathcal{Y}}$ in the input and $f_{\omega_\mathcal{X}}$ in the target domain. Perceptual similarity is compared between each generator input and its reconstruction. The right plot indicates that during inference, only $G_{\theta_\mathcal{Y}}$ has to be deployed to synthesize new depth profiles. RGB images and ground truth depth images were taken from the Texas-3DFRD [12]

structure by taking into account the similarity between the input and the reconstruction of the synthesized depth profile. Obviously, calculation of a reconstruction error requires an opposite generator function $G_{\theta_\mathcal{X}} : \mathcal{Y} \to \mathcal{X}$ to assimilate real RGB distribution $P_\mathcal{X}$ as well as the corresponding distance network $f_{\omega_\mathcal{X}} : \mathcal{X} \to \mathbb{R}$. Both have to be optimized simultaneously to the RGB-to-depth direction. The reconstruction error is commonly evaluated by assessing similarity between x and $G_{\theta_\mathcal{X}}(G_{\theta_\mathcal{Y}}(x))$ as well as similarity between y and $G_{\theta_\mathcal{Y}}(G_{\theta_\mathcal{X}}(y))$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. In the setting of style transfer and cycle-consistent GANs [20], a pixel-wise distance function on image space is considered, where the mean absolute error (MAE) or the mean squared error (MSE) is the common choice.

The use of a contrary generator $G_{\theta_\mathcal{X}}$ can be viewed as a type of regularization since it prevents mode collapse, i.e., generator outputs remain dependent on the inputs. Deployment of the cycle-consistency approach [20], where reconstruction error is measured in image space, assumes no information loss during the modality transition. This corresponds to the applications of summer-to-winter landscape or photograph-to-Monet painting transition. Determining $G_{\theta_\mathcal{Y}}$ and $G_{\theta_\mathcal{X}}$ is an ill-posed problem since a single depth profile may be generated by an infinite number of distinct RGB images and vice versa [42]. For example, during RGB-to-depth transition of human faces, information on image brightness, light source or the subject's skin color is lost. As a consequence, the contrary depth-to-RGB generator needed for regularization has to synthesize the lost properties of the image. Both generators $G_{\theta_\mathcal{Y}}$ and $G_{\theta_\mathcal{X}}$ may be penalized if the skin color or the brightness of the reconstruction is changed even though $G_{\theta_\mathcal{X}}$ did exactly what we expected it to do, i.e., synthesize a face that is related to the input's depth profile.

Adapting the idea of [43], we propose a perceptual reconstruction loss, i.e., instead of computing a reconstruction error in image space, we consider certain image features of the reconstruction. Typical perceptual similarity metrics extract features by propagating the images (to be compared) through an auxiliary network that is usually pretrained on a large image classification task [43–45]. Nevertheless, we expect our feature extractor to be perfectly tailored to our data and not determined by an additional network pretrained on a very general classification task [44] that may not even cover our type of data. Therefore, we enforce the reconstruction consistency on the image space by using the MAE loss on feature vectors extracted by $\phi_\mathcal{X}(\cdot) := f^l_{\omega_\mathcal{X}}(\cdot)$, which corresponds to the l-th layer of the RGB critic (cf. Algorithm 1). Analogously, we define the feature extractor on depth space by $\phi_\mathcal{Y}(\cdot) := f^l_{\omega_\mathcal{Y}}(\cdot)$, which corresponds to the l-th layer of the depth critic. Although we are aware that feature extractor weights are adjusted with each update of critic weights $\omega_\mathcal{X}, \omega_\mathcal{Y}$, we assume that, at least at a later stage of training, $\phi_\mathcal{X}$ and $\phi_\mathcal{Y}$ have learned good and stable features on the image and depth domain. This yields the following empirical reconstruction risk:

$$\begin{aligned}\mathcal{R}_{\text{rec}}(\theta_\mathcal{X}, \theta_\mathcal{Y}, \phi_\mathcal{X}, \phi_\mathcal{Y}, \mathbf{x}, \mathbf{y}) := \frac{1}{b}\sum_{n=1}^{b} \text{MAE} \\ \left[\phi_\mathcal{X}\left(G_{\theta_\mathcal{X}}\left(G_{\theta_\mathcal{Y}}(x_n)\right)\right), \phi_\mathcal{X}(x_n)\right] \\ + \frac{1}{b}\sum_{n=1}^{b} \text{MAE}\left[\phi_\mathcal{Y}\left(G_{\theta_\mathcal{Y}}\left(G_{\theta_\mathcal{X}}(y_n)\right)\right), \phi_\mathcal{Y}(y_n)\right].\end{aligned} \tag{4}$$

In our implementation, we set $l{:=}L-2$ for a critic with L layers, i.e., we use the second-to-last layer of the critic.
A good reconstruction term must still be found for the start of training when the critic features are not yet sufficiently reliable. At first, it is desirable to guide the framework to preserve structural similarity during RGB-to-depth and depth-to-RGB transition. Therefore, we propose to compare the input and its reconstruction in the image space while automatically removing the brightness, illumination and color of the RGB images beforehand. This can be ensured by applying the following steps:

1. Convert the image to grayscale by applying the function $g\colon [0,255]^{d_1\times d_2\times 3} \to \mathbb{R}^{d_1\times d_2},\quad x \mapsto \frac{0.299}{255}\cdot x_{(,,0)} + \frac{0.587}{255}\cdot x_{(,,1)} + \frac{0.144}{255}\cdot x_{(,,2)}$, where $(,,i)$ denotes the i-th color channel for $i=0,1,2$.
2. Enhance the brightness of the grayscale image using an automated gamma correction based on the image brightness [46], i.e. take the grayscale image x_{gr} to the power of $\Gamma(x_{\mathrm{gr}}){:=}{-}0.3\cdot 2.303/\ln \overline{x_{\mathrm{gr}}}$, where $\overline{x_{\mathrm{gr}}}$ denotes the average of the gray values.
3. Convolve the enhanced image with a high-pass filter h in order to dim the lighting source and color information (cf. Fig. 3). The high-pass filter may be applied in Fourier domain, i.e., the 2D Fourier transform is multiplied by a Gaussian high-pass filter matrix H^σ defined by $H^\sigma_{i,j}{:=}1-\exp\left(\left\|(i,j)-(\frac{d_1}{2},\frac{d_2}{2})\right\|_2^2/(2\sigma^2)\right)$ for $i=1,\dots,d_1$ and $j=1,\dots,d_2$. In our implementation, $\sigma=4$ yielded satisfactory results for all tasks.

This yields the updated empirical reconstruction risk:

$$
\begin{aligned}
&\mathcal{R}_{\mathrm{rec}}(\theta_{\mathcal{X}},\theta_{\mathcal{Y}},\phi_{\mathcal{X}},\phi_{\mathcal{Y}},\gamma,\mathbf{x},\mathbf{y})\\
&:=\gamma\cdot\frac{1}{b}\sum_{n=1}^{b}\mathrm{MAE}\big[\phi_{\mathcal{X}}\big(G_{\theta_{\mathcal{X}}}\big(G_{\theta_{\mathcal{Y}}}(x_n)\big)\big),\phi_{\mathcal{X}}(x_n)\big]\\
&\quad+\gamma\cdot\frac{1}{b}\sum_{n=1}^{b}\mathrm{MAE}\big[\phi_{\mathcal{Y}}\big(G_{\theta_{\mathcal{Y}}}\big(G_{\theta_{\mathcal{X}}}(y_n)\big)\big),\phi_{\mathcal{Y}}(y_n)\big]\\
&\quad+(1-\gamma)\cdot\frac{1}{b}\sum_{n=1}^{b}\mathrm{MAE}\big[\psi\big(G_{\theta_{\mathcal{Y}}}\big(G_{\theta_{\mathcal{X}}}(x_n)\big)\big),\psi(x_n)\big]\\
&\quad+(1-\gamma)\cdot\frac{1}{b}\sum_{n=1}^{b}\mathrm{MAE}\big[G_{\theta_{\mathcal{Y}}}\big(G_{\theta_{\mathcal{X}}}(y_n)\big),y_n\big],
\end{aligned}
\tag{5}
$$

where $\psi(\cdot){:=}h*g(\cdot)^{\Gamma(g(\cdot))}$ and γ is gradually increased from 0 to 1 during training to control feature extractor reliability. In the far right column in Fig. 3, we may observe the strong effect of operator ψ. For the face sample, the face shape and the positions of the nose and the eyes are very clear, at the same time the low image brightness and the exposure direction are resolved. The main edges of the cylinder liner surfaces are clearly identifiable, whereas the different brown levels and illumination inconsistencies of the input are no longer visible.

Using the previously discussed risk functions $\mathcal{R}_{\mathrm{cri}}$ (2), $\mathcal{R}_{\mathrm{adv}}$ (3) and $\mathcal{R}_{\mathrm{rec}}$ (5), Algorithm 1 summarizes the proposed architecture for fully unsupervised single-view depth estimation. Implementation of the proposed framework is publicly available on https://github.com/anger-man/unsupervised-depth-estimation.

Algorithm 1 Proposed Framework

Require: α_f critic learning rate; α_G generator learning rate; p gradient penalty; n_f number of critic iterations; n_G number of generator updates; b minibatch size; λ_{rec} reconstruction loss weight
Require: $\omega_{\mathcal{Y}},\omega_{\mathcal{X}}$ initial critic weights; $\theta_{\mathcal{Y}},\theta_{\mathcal{X}}$ initial generator weights; $\gamma=0$
for $k=1,\dots,n_G$ **do**
  for $i=1,\dots,n_f$ **do**
    Sample $\mathbf{x}=\{x_n\}_{n=1}^b\subset X$ and $\mathbf{y}=\{y_n\}_{n=1}^b\subset Y$
    $\{\tilde{y}_n\}_{n=1}^b \leftarrow \left\{\epsilon_n\cdot G_{\theta_{\mathcal{Y}}}(x_n)+(1-\epsilon_n)\cdot y_n,\ \epsilon_n\sim\mathcal{U}[0,1]\right\}_{n=1}^b$
    $\{\tilde{x}_n\}_{n=1}^b \leftarrow \left\{\epsilon_n\cdot G_{\theta_{\mathcal{X}}}(y_n)+(1-\epsilon_n)\cdot x_n,\ \epsilon_n\sim\mathcal{U}[0,1]\right\}_{n=1}^b$
    $\partial_{\mathcal{Y}} \leftarrow \nabla_{\omega_{\mathcal{Y}}}\mathcal{R}_{\mathrm{cri}}(\omega_{\mathcal{Y}},\theta_{\mathcal{Y}},p,\mathbf{y},\mathbf{x})$
    $\partial_{\mathcal{X}} \leftarrow \nabla_{\omega_{\mathcal{X}}}\mathcal{R}_{\mathrm{cri}}(\omega_{\mathcal{X}},\theta_{\mathcal{X}},p,\mathbf{x},\mathbf{y})$
    $\omega_{\mathcal{Y}} \leftarrow \mathrm{Adam}(\omega_{\mathcal{Y}},\partial_{\mathcal{Y}},\alpha_f,\beta_1=0,\beta_2=0.9)$
    $\omega_{\mathcal{X}} \leftarrow \mathrm{Adam}(\omega_{\mathcal{X}},\partial_{\mathcal{X}},\alpha_f,\beta_1=0,\beta_2=0.9)$
  end for
  Sample $\mathbf{x}=\{x_n\}_{n=1}^b\subset X$ and $\mathbf{y}=\{y_n\}_{n=1}^b\subset Y$; set $\phi_{\mathcal{Y}},\phi_{\mathcal{X}}$ to l-th layer of $f_{\omega_{\mathcal{Y}}},f_{\omega_{\mathcal{X}}}$
  $\partial_{\mathcal{Y}} \leftarrow \nabla_{\theta_{\mathcal{Y}}}\mathcal{R}_{\mathrm{adv}}(\theta_{\mathcal{Y}},\omega_{\mathcal{Y}},\mathbf{x})+\lambda_{\mathrm{rec}}\cdot\nabla_{\theta_{\mathcal{Y}}}\mathcal{R}_{\mathrm{rec}}(\theta_{\mathcal{X}},\theta_{\mathcal{Y}},\phi_{\mathcal{X}},\phi_{\mathcal{Y}},\gamma,\mathbf{x},\mathbf{y})$
  $\partial_{\mathcal{X}} \leftarrow \nabla_{\theta_{\mathcal{X}}}\mathcal{R}_{\mathrm{adv}}(\theta_{\mathcal{X}},\omega_{\mathcal{X}},\mathbf{y})+\lambda_{\mathrm{rec}}\cdot\nabla_{\theta_{\mathcal{X}}}\mathcal{R}_{\mathrm{rec}}(\theta_{\mathcal{X}},\theta_{\mathcal{Y}},\phi_{\mathcal{X}},\phi_{\mathcal{Y}},\gamma,\mathbf{x},\mathbf{y})$
  $\theta_{\mathcal{Y}} \leftarrow \mathrm{Adam}(\theta_{\mathcal{Y}},\partial_D,\alpha_G,\beta_1=0,\beta_2=0.9)$
  $\theta_{\mathcal{X}} \leftarrow \mathrm{Adam}(\theta_{\mathcal{X}},\partial_C,\alpha_G,\beta_1=0,\beta_2=0.9)$
  $\gamma \leftarrow \frac{k}{n_G}$
end for

3.3 Network implementation

As critical as the loss function design of an unsupervised method is the choice of an appropriate architecture for the critic and the generator network. A decoder for the critic is built following the PacthGAN critic that was initially proposed in [32] with nearly 15.7×10^6 parameters. The PatchGAN architecture has been found to perform quite stably over a variety of different generative task and is part of many state-of-the-art architectures for image generation [20, 24, 47]. The generator is a ResNet18 [48] with a depth-specific upsampling part taken from [17] (19.8×10^6 parameters). Detailed information on critic and generator implementations is provided in the supplementary.

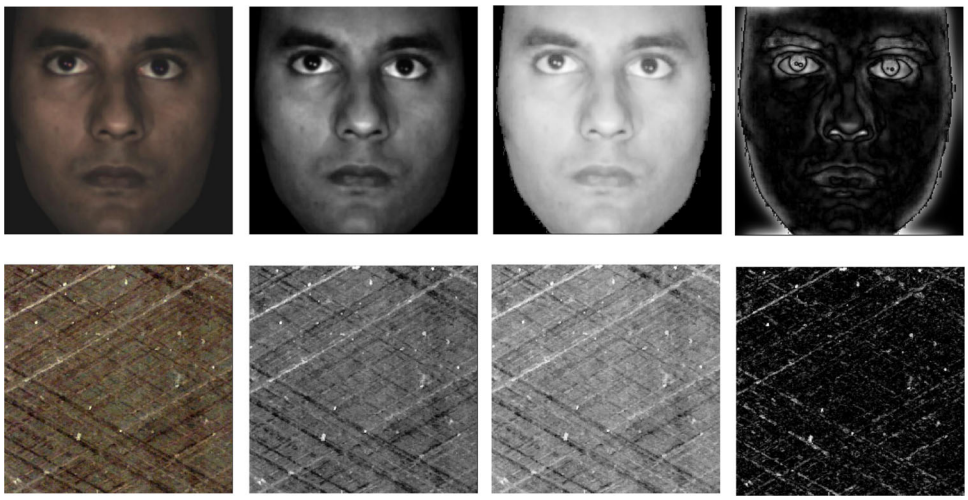

Fig. 3 The first column visualizes the RGB samples and the second column the grayscale versions. The third column contains the gamma corrected counterparts, where the contrast in lower gray levels is enhanced for dark images in particular. The last column illustrates the application of the high-pass filter

4 Experiments and discussion

The framework proposed in Algorithm 1 is implemented with the publicly available TensorFlow framework [49]. The applications are inner surface depth estimation of cylinder liners, face depth estimation based on the Texas-3DFRD [12] and body depth synthesis using the SURREAL dataset [9]. In this section, we benchmark the proposed framework on each dataset and separately present the results, followed by a discussion at the end. As discussed in the introduction, the methods used for comparison are cycleGAN [20], gcGAN [47] and CUT [24]. For CUT, we use the public github repository. [1] The benchmark methods cycleGAN and gcGAN use the same critic and generator implementations as the method proposed in this study (cf. Sect. 3.3). For cycleGAN, we remove the novel perceptual loss and hand-crafted image filters from our method and replace them with MAE reconstruction loss. For gcGAN, the contrary generator is removed and up-down flip is employed as the geometric constraint.

An ablation study is conducted in order to highlight the impact and necessity of the novel hand-crafted filters and the perceptual reconstruction loss proposed here. More precisely, we set $\psi \triangleq \mathrm{Id}$ in (5) in order to avoid the hand-crafted filters, i.e., the reconstruction loss in the RGB domain is determined based on the MAE. Furthermore, we set $\gamma \triangleq 0$ in (5) for the entire training process to study network behavior without perceptual reconstruction in both domains. The experiments without hand-crafted filters (w/o ψ) and without perceptual reconstruction (w/o ϕ) are performed for the test cases of surface depth and face depth estimation (cf. Tables 1 and 2).

In our implementation, we set the number of generator updates n_G to 10k, the minibatch size b to 8 and the penalty term p to 100. The number of critic iterations n_f is initially established to be 24 to ensure a good approximation of the Wasserstein-1 distance in the beginning. After 1000 generator updates, it is halved to speed up training. Furthermore, we set α_f to 5×10^{-5} and α_G to 1×10^{-4}. The influence of the reconstruction term λ_{rec} is found for each dataset and method individually by a parameter grid search.

[1] https://github.com/taesungp/contrastive-unpaired-translation.

Table 1 Unsup. surface depth estimation: the reported metrics are RMSE and MAE of the ground truth and the synthesized depth and are evaluated on unseen data (smaller is better)

Method	Two-sided	λ_{rec}	RMSE $\pm$ std (μm)	MAE $\pm$ std (μm)
Ours	✓	10	**0.751** $\pm$ 0.195	**0.533** $\pm$ 0.144
Ours w/o ϕ	✓	10	0.788 $\pm$ 0.226	0.568 $\pm$ 0.171
Ours w/o ψ	✓	10	0.820 $\pm$ 0.230	0.586 $\pm$ 0.174
cycleGAN	✓	2	0.833 $\pm$ 0.175	0.600 $\pm$ 0.132
gcGAN	**x**	1	0.777 $\pm$ 0.196	0.555 $\pm$ 0.145
CUT	**x**	10	1.434 $\pm$ 0.402	1.074 $\pm$ 0.326

When comparing the individual methods, the best achieved values with regard to the validation metrics are marked in bold

Table 2 Unsup. face depth estimation: the reported metrics are RMSE and MAE of the ground truth and the synthesized depth and are evaluated on unseen data (smaller is better)

Method	Two-sided	λ_{rec}	RMSE $\pm$ std	MAE $\pm$ std
Ours	✓	10	**0.064** $\pm$ 0.022	**0.046** $\pm$ 0.017
Ours w/o ϕ	✓	10	0.072 $\pm$ 0.028	0.053 $\pm$ 0.024
Ours w/o ψ	✓	10	0.089 $\pm$ 0.037	0.070 $\pm$ 0.032
cycleGAN	✓	1	0.105 $\pm$ 0.049	0.073 $\pm$ 0.033
gcGAN	**x**	0.3	0.078 $\pm$ 0.039	0.058 $\pm$ 0.034
CUT	**x**	10	0.094 $\pm$ 0.039	0.081 $\pm$ 0.042

When comparing the individual methods, the best achieved values with regard to the validation metrics are marked in bold

Table 3 Unsup. body depth estimation: the reported metrics are RMSE and MAE of the ground truth and the synthesized depth and are evaluated on unseen data (smaller is better)

Method	Two-sided	λ_{rec}	RMSE $\pm$ std (m)	MAE $\pm$ std (m)
Proposed	✓	1	**0.080** $\pm$ 0.033	**0.022** $\pm$ 0.020
cycleGAN	✓	1	0.091 $\pm$ 0.035	0.033 $\pm$ 0.019
gcGAN	x	1	0.095 $\pm$ 0.036	0.030 $\pm$ 0.021
CUT	x	1	0.183 $\pm$ 0.021	0.071 $\pm$ 0.016

When comparing the individual methods, the best achieved values with regard to the validation metrics are marked in bold

4.1 Surface depth

This study uses the same database initially proposed in [4] for depth estimation of inner cylinder liner surfaces of large internal combustion engines.

Depth measurements cover a spatial region of 1.9 mm × 1.9 mm, have a dimension of approximately 4000×4000 pixels and are acquired using a resource-intensive logistic chain as discussed in the introduction. The profiles denote relative depth with respect to the core area of the surface on a µm scale. The RGB data is taken from the same cylinder surfaces with a simple handheld microscope. The RGB measurements cover a region of 4.2 mm × 4.2 mm and have a resolution of nearly 1024×1024 pixels. The smaller image area of the depth measurements is not registered in the larger RGB area, but the RGB instances are randomly cropped to 1.9 mm × 1.9 mm to ensure the same spatial size between RGB and depth data. 592 random samples are obtained from each image domain. The RGB and depth data is then augmented separately to nearly 7000 samples via random cropping, flipping and gamma correction [46]. To make computation feasible on a *NVIDIA GeForce RTX 2080* GPU, each sample is resized to a dimension of 256×256 pixels. In order to assess the visual quality between two completely unaligned domains, we also generated depth profiles of 211 additional surface areas and registered them with great effort using shear transformations and a mutual information criterion. These evaluation samples are not included in the training database. During optimization, RGB images and depth profiles are scaled from $[0, 255]$ to $[-1, 1]$ and from $[-5, 5]$ to $[-1, 1]$, respectively, whereas evaluation metrics (RMSE and MAE) are calculated on the original depth scale in µm.

4.2 Face depth

The Texas-3DFRD [12] consists of 118 individuals and a variety of facial expressions and corresponding depth profiles are available for each of them. Depth pixels represent absolute depth and their values are in $[0, 1]$ where 1 represents the near clipping plane while 0 denotes the background. We randomly select 16 individuals as evaluation data and use the remaining samples as training data. For unsupervised training, we randomly select 50% of the training individuals for the input domain and use the depth images of the remaining 50% for the target domain. We resize all RGB frames and depth profiles to a dimension of 256 × 256 pixels. Data is augmented via flipping, histogram equalization and Gaussian blur to nearly 6300 samples per modality. During optimization, RGB images are scaled from $[0, 255]$ to $[-1, 1]$ and depth profiles are scaled from $[0, 1]$ to $[-1, 1]$, whereas the evaluation metrics RMSE and MAE are computed on the original depth scale.

More experiments on unsupervised facial depth synthesis on the Bosphorus-3DFA [11], the CelebAMask-HQ [23] and qualitative comparison to Wu et al. [26] are presented in the supplementary.

4.3 Body depth

The SURREAL dataset [9] consists of nearly 68k video clips that show 145 different synthetic subjects performing various actions. The clips consist of 100 RGB frames with perfectly aligned depth profiles that denote real-world camera distance. We use the same train/test split as Varol et al. [9], i.e., we remove nearly 12.5 k clips and use the middle frame of each 100-frame clip for evaluation. For the remaining clips, an amount of 2500 clips is randomly selected for training. We choose 20 RGB and 20 depth frames per clip ensuring that RGB and depth frames are disjointed in order to mimic an application without any accurately aligned RGB-depth pairs. This results in approximately 50 k samples per modality. We strictly follow the preprocessing pipeline of Varol et al. [9], cropping each frame to the human bounding box and resizing/padding images to a dimension of 256 × 256 pixels.

In addition, for each image, we subtract the median of depth values to fit the depth images into the range ±0.4725 m, where values less or equal −0.4725 denote background. During optimization, RGB images are scaled from $[0, 255]$ to $[-1, 1]$ and depth profiles are scaled from $[-0.4725, 0.4725]$ to $[-1, 1]$, whereas evaluation metrics RMSE and MAE are computed on the original depth scale in meters.

4.4 Discussion

Quantitative evaluation on unseen test data in Tables 1, 2 and 3 confirms superiority of the proposed method compared to other state-of-the-art modality transfer methods. In particular, the CUT method is not suitable for the depth estimation of planar surfaces and human bodies. Obviously, usage of a novel perceptual reconstruction term in combination with hand-crafted image filters is able to overcome the shortcomings of a standard cycle-consistency constraint as explained in Sect. 3.2 and improves depth accuracy significantly. Considering the industrial application, Fig. 4 and Fig. 5 indicates

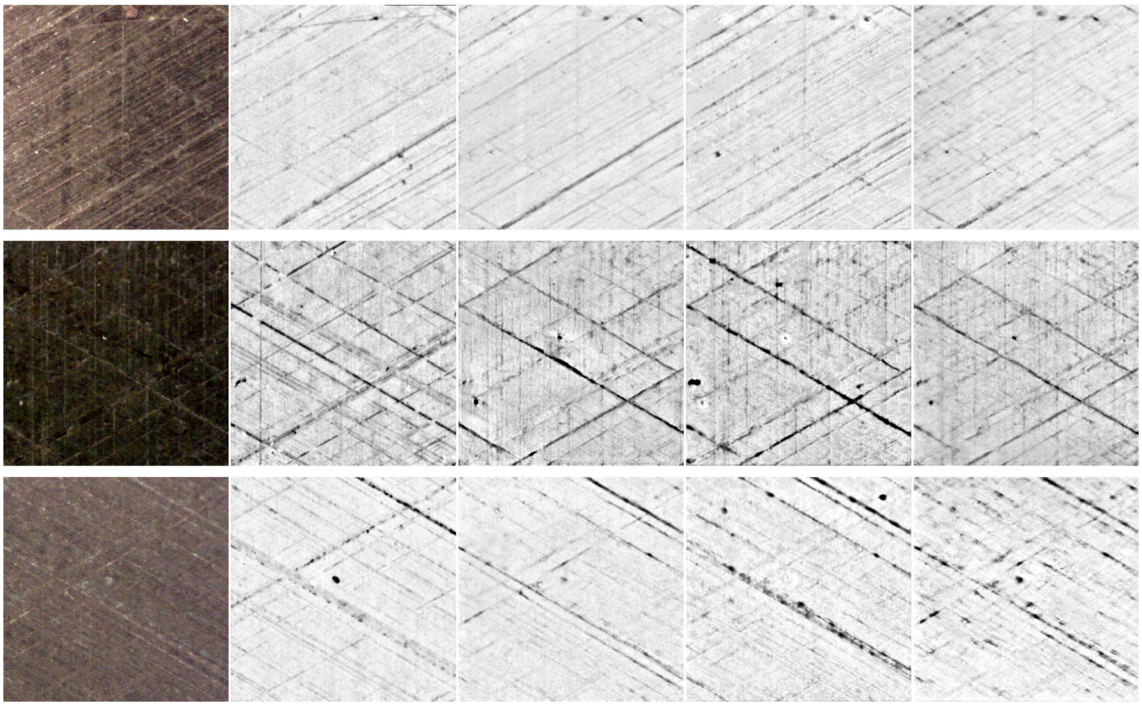

Fig. 4 From left to right: surface RGB input, ground truth and profiles predicted by our method, gcGAN and cycleGAN

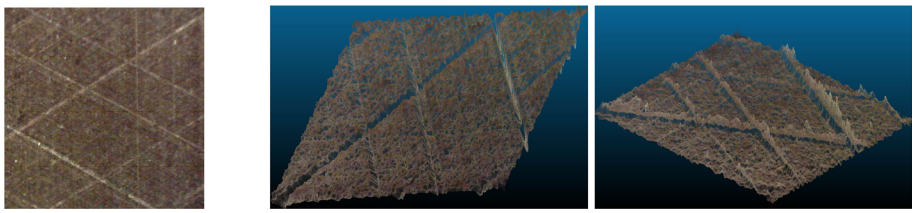

Fig. 5 An instant 3D model generated by our proposed framework provides valuable information on the liner surface condition

that we have been able to synthesize realistic surface depth profiles with an RMSE of 0.751 μm compared to the registered ground truth. In Fig. 6, we observe that predictions coming from our method seem most similar to the ground truth, while the results of cycleGAN and CUT do not correctly reproduce the contours of the input. Plausibility of our depth predictions is also confirmed by the instant 3D model in Fig. 7. In Fig. 8, it can be seen that the CUT benchmark completely fails on the SURREAL dataset, which can possibly be attributed to the fact that here, in parallel to the depth estimation, the body must also be segmented.

Although the proposed method was initially motivated by cycleGAN [20], it is important to point out that replacement of the standard cycle-consistency term with perceptual losses and usage of appropriate hand-crafted filters in image space is a novel idea that overcomes significant shortcomings of the standard cycleGAN architecture in depth estimation that are thoroughly discussed in the paper. For depth synthesis of surfaces, faces and human bodies, the RMSE decreases (compared to a standard cycleGAN) about 9.8%, 39.1% and 12.1%, respectively. Tables 1 and 2 show how the removal of the perceptual reconstruction loss (w/o ϕ) and the hand-crafted filters (w/o ψ) reduces the accuracy of the proposed method. However, the use of perceptual-based reconstruction and the inclusion of hand-crafted filters each outperform the cycleGAN benchmark, with the combination of the two techniques providing the best performance in terms of evaluation metrics. The proposed method has been mainly developed to find a solution to the problem of depth synthesis of planar cylinder liner surfaces. The results confirm that the framework not only succeeds on the cylinder surface task but also significantly improves performance in the field of face and whole body depth synthesis compared to state-of-the-art modality transfer methods.

All three prototypical studies of single-shot depth prediction have in common that the color of the objects in the RGB instances has nearly no effect on the depth. This was the main motivation for the hand-crafted filters that convert RGB instances to gray values and remove low-frequency components. However, the motivation for these filters does not apply to all depth estimation problems. Indeed, there are examples where the color of the RGB instance could also give an indi-

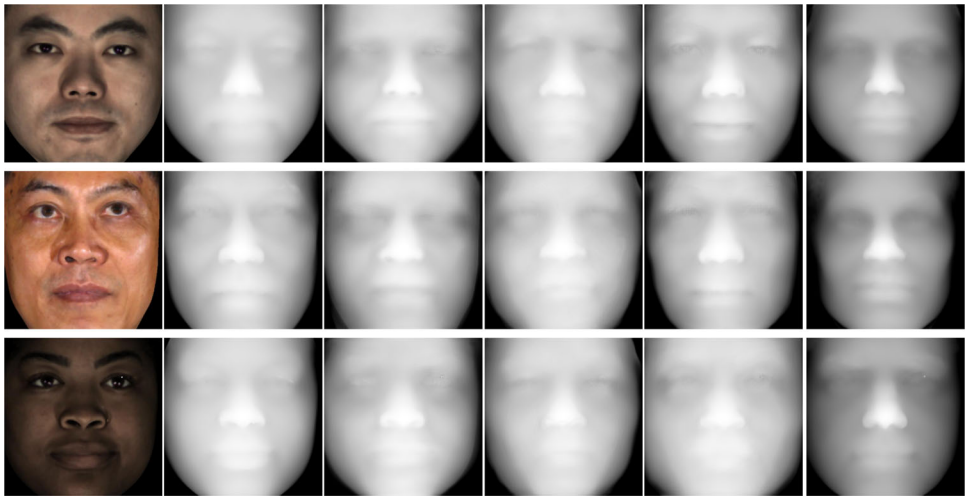

Fig. 6 From left to right: Face RGB input, ground truth and profiles predicted by our method, gcGAN, cycleGAN and CUT

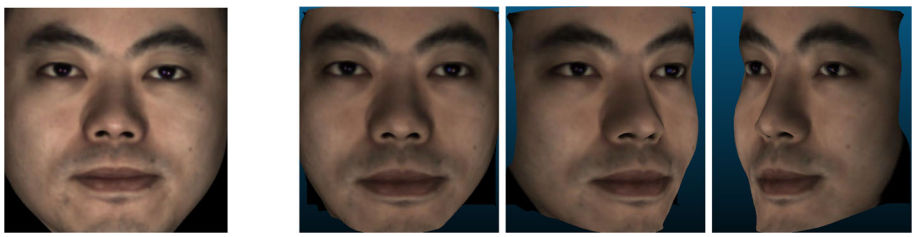

Fig. 7 An example of viewpoint augmentation using a 3D face model instantly generated by our proposed framework

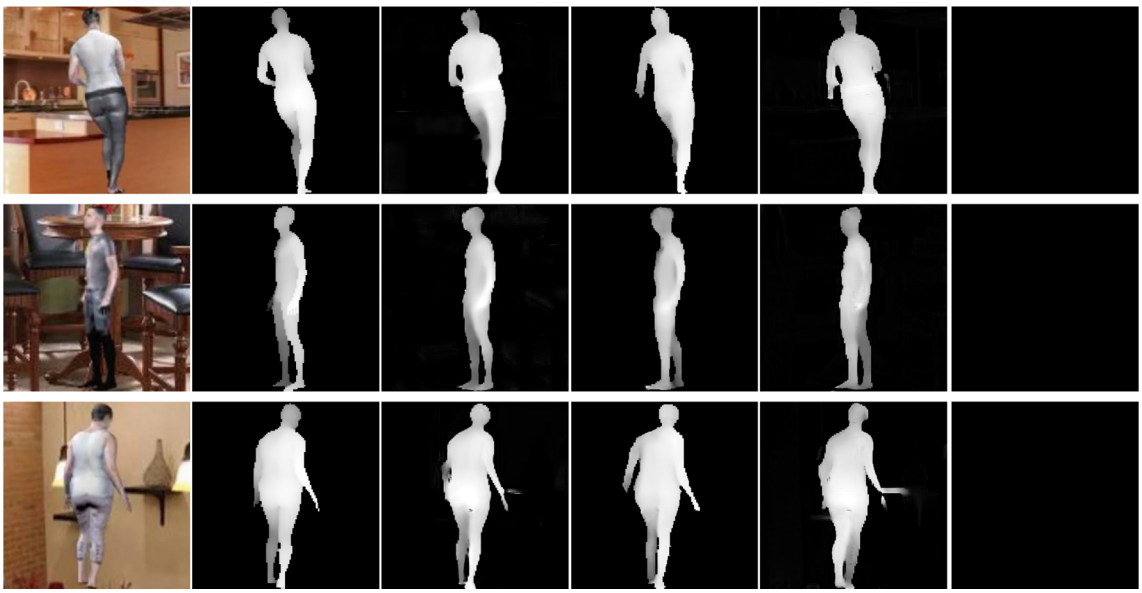

Fig. 8 From left to right: Body RGB input, ground truth and profiles predicted by the proposed method, gcGAN, cycleGAN and CUT

cation of the depth of the observed scene. An example would be depth estimation from satellite images, i.e., modeling altitude from aerial imagery data. In such cases, the structure of the hand-crafted filters must be reconsidered and adjusted accordingly.

5 Conclusion

This paper proposes a framework for fully unsupervised single-shot depth estimation from monocular RGB images based on the Wasserstein-1 distance, a novel perceptual reconstruction loss and hand-crafted image filters. The model is comprehensively evaluated on differing depth synthesis tasks without using pairwise RGB and depth data during

training. The approach provides a reasonable solution for estimating the relative depth of cylinder liner surfaces when generation of paired data is technically not feasible. Moreover, the proposed algorithm also shows promising results when applied to the task of absolute depth estimation of human bodies and faces, thereby proving that it may be generalized to other real-life tasks.

However, one disadvantage of the perceptual reconstruction approach is that four neural networks must be fitted in parallel.

Future work will therefore include the development of one-sided depth synthesis models in an unsupervised manner as well as the application of our approach to other modality transfer tasks.

Acknowledgements The authors would like to acknowledge the financial support of the "COMET - Competence Centres for Excellent Technologies" Programme of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK) and the Federal Ministry of Labour and Economy (BMAW) and the Provinces of Styria, Tyrol and Vienna for the COMET Centre (K1) LEC EvoLET. The COMET Programme is managed by the Austrian Research Promotion Agency (FFG) (Grant No. 865843).

Funding Open access funding provided by University of Innsbruck and Medical University of Innsbruck.

Appendix A: 3D databases: an overview

Single-shot depth estimation has become increasingly popular over the last decade of deep learning. The first deep learning solutions for depth synthesis were motivated by the development of autonomous driving and localization systems and therefore were initially designed to automatically determine the depth of indoor or outdoor scenes [14–17, 38–40]. Deep convolutional neural networks, trained on large-scale and extensive data sets such as KITTI [2] or NYU Depth Dataset v2 [1] achieved state-of-the-art results. The outdoor video clips of the KITTI dataset can be used for various subtasks in computer vision such as optical flow, object detection, semantic segmentation and depth [3]. Each video sequence of the KITTI dataset consists of stereo image pairs with aligned depth images (LIDAR), which renders the database a common benchmark for unsupervised or self-supervised depth estimation tasks [16, 17, 38]. The NYU Depth Dataset v2 focuses on monocular sequences of indoor environments, where depth counterparts are obtained with a high quality RGB-D camera. Therefore, this dataset is considered a primary benchmark in supervised monocular depth estimation [14, 15].

With the advent of virtual and augmented reality applications, single-image pose estimation and 3D reconstruction of human bodies or body parts received a great amount of attention in the research field of computer vision [18]. 3D information on human faces provides additional benefits for face recognition or detection systems [6]. The Texas-3DFRD [12] and the Bosphorus-3DFA [11] are known representatives of paired face RGB-depth data of high quality and include a variety of head poses and emotional expressions. Both databases provide facial landmarks for additional face expression analysis, but with approximately 100 different individuals each, the sets are rather small. A larger number of facial depth models can be derived from 3D synthetic data of human faces as in [7, 50]. Leveraging the task to whole body depth estimation is challenging due to the fact that RGB-depth pairs of real individuals are not abundant in many datasets. A small dataset of 25 video clips for detailed human depth estimation is proposed in [10], while a depth dataset of 10 sequences recorded from different viewpoints is published in [8]. The Human3.6M dataset [13] contains high-resolution depth data from 11 individuals acting in varying scenarios. Varol et al. [9] propose using the approximately 68k video clips of synthetic humans in the large-scale SURREAL dataset for supervised training of human body depth and segmentation models.

Appendix B: Network details

In Tables 4, 5 and 6, **k** denotes the kernel size, **s** the stride, and **channels** the number of layer output channels. **Input** corresponds to the input of each layer. Network input and output are denoted by $\mathcal{I}$ and $\mathcal{O}$, respectively, where for a generator network the output channel size equals 1 (RGB-to-depth) or 3 (depth-to-RGB).

Table 4 ResNet18 generator

Name	Type	k	s	chns	Input	Activ.
con1	conv-norm	7	2	64	$\mathcal{I}$	ReLU
max1	maxpool3x3		2	64	con1	
res1	res-block	3	1	64	max1	ReLU
res2	res-block	3	1	64	res1	ReLU
res3	res-block	3	2	128	res2	ReLU
res4	res-block	3	1	128	res3	ReLU
res5	res-block	3	2	256	res4	ReLU
res6	res-block	3	1	256	res5	ReLU
res7	res-block	3	2	512	res6	ReLU
res8	res-block	3	1	512	res7	ReLU
ups1	upsampling		2	512	res8	
con2	conv-norm	3	1	512	ups1	ELU
cct1	concatenate			768	con2,res6	
con3	conv-norm	3	1	512	cct1	ELU
ups2	upsampling		2	512	con3	
con4	conv-norm	3	1	256	ups2	ELU
cct2	concatenate			384	con4,res4	
con5	conv-norm	3	1	256	cct2	ELU
ups3	upsampling		2	256	con5	
con6	conv-norm	3	1	128	ups3	ELU
cct3	concatenate			192	con6,res2	
con7	conv-norm	3	1	128	cct3	ELU
ups4	upsampling		2	128	con7	
con8	conv-norm	3	1	64	ups4	ELU
cct4	concatenate			128	con8,con1	
con9	conv-norm	3	1	64	cct4	ELU
ups5	upsampling		2	64	con9	
con10	conv-norm	3	1	32	ups5	ELU
con11	conv-norm	3	1	32	con10	ELU
$\mathcal{O}$	convolution	3	1	3/1	con11	tanh

The encoder is quite similar to the illustrated architecture in [48]. The decoder architecture is a slightly modified version of [17]. For upsampling, nearest neighbor method is used. Convolution layers followed by an instance normalization are denoted by conv-norm

Table 5 PatchGAN critic

Name	Type	k	s	chns	Input	Activ.
con1	convolution	4	1	16	$\mathcal{I}$	LReLU
con2	convolution	4	1	16	con1	LReLU
con3	convolution	4	2	32	con2	LReLU
con4	convolution	4	1	32	con3	LReLU
con5	convolution	4	2	64	con4	LReLU
con6	convolution	4	1	64	con5	LReLU
con7	convolution	4	2	128	con6	LReLU
con8	convolution	4	1	128	con7	LReLU
con9	convolution	4	2	256	con8	LReLU
con10	convolution	4	1	256	con9	LReLU
con11	convolution	4	2	512	con10	LReLU
con12	convolution	4	1	512	con11	LReLU
$\mathcal{O}$	convolution	4	1	1	con12	linear

LReLU denotes the Leaky ReLU activation function with slope parameter 0.2

Table 6 Residual block

Name	Type	k	s	chns	Input	Activ.
con1	conv-norm	k	s	c	$\mathcal{I}$	ReLU
con2	conv-norm	k	s	c	con1	
skip	conv-norm	1	s	c	$\mathcal{I}$	
add	addition			c	con2,skip	
$\mathcal{O}$	activation			c	add	ReLU

A residual block (res-block) with kernel size k, stride s and channel size c is implemented as follows

Appendix C: Facial depth estimation on bosphorus-3DFA and CelebAMask-HQ

Section 4.2 demonstrates the plausibility of our proposed framework for fully unsupervised facial depth estimation using the small Texas-3DFRD [12]. Obviously, the shooting position of the portrayed faces is always constant. The data set consists exclusively of frontal views, the illumination direction is consistent, and all images are individually cropped to the facial region. However, the goal of this section is to train a model that is capable of generating depth profiles from arbitrary portrait images that are at least sufficient for reasonable viewpoint augmentation. To accomplish this, we make use of the following two data sets: the Bosphorus Database for 3D Face Analysis (Bosphorus-3DFA) [11] and the CelebAMask-HQ [23] that records face portraits.

The Bosphorus-3DFA consists of 105 individuals, where for each person, in contrast to the Texas-3DFRD, varying poses, different head rotations and occlusions (e.g. eyeglasses, long hair) are available. Pixel-aligned depth samples represent absolute depth and are preprocessed to the range [0, 1].

Analogously to Sect. 4.2, we resize all RGB frames and depth profiles to a dimension of 256×256 and conduct data augmentation via random cropping. This results into 11k samples per modality. Although this database now contains different positions and face expressions, the decisive disadvantage is that all images were taken with constant lighting and with the same background (cf. Fig. 9). Therefore, we add the CelebAMask-HQ to our experiment.

The CelebAMask-HQ is a large-scale facial portrait dataset with high-resolution face images of 30k celebrities selected from the CelebA dataset [51]. Each sample is provided with a segmentation mask of face attributes, and therefore this database is used to train and evaluate face

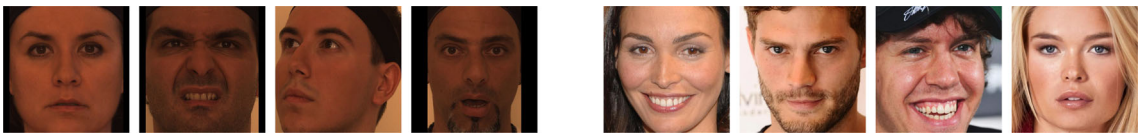

Fig. 9 Left: RGB samples of the Bosphorus-3DFA [11]. Right: Samples of the CelebAMask-HQ [23]

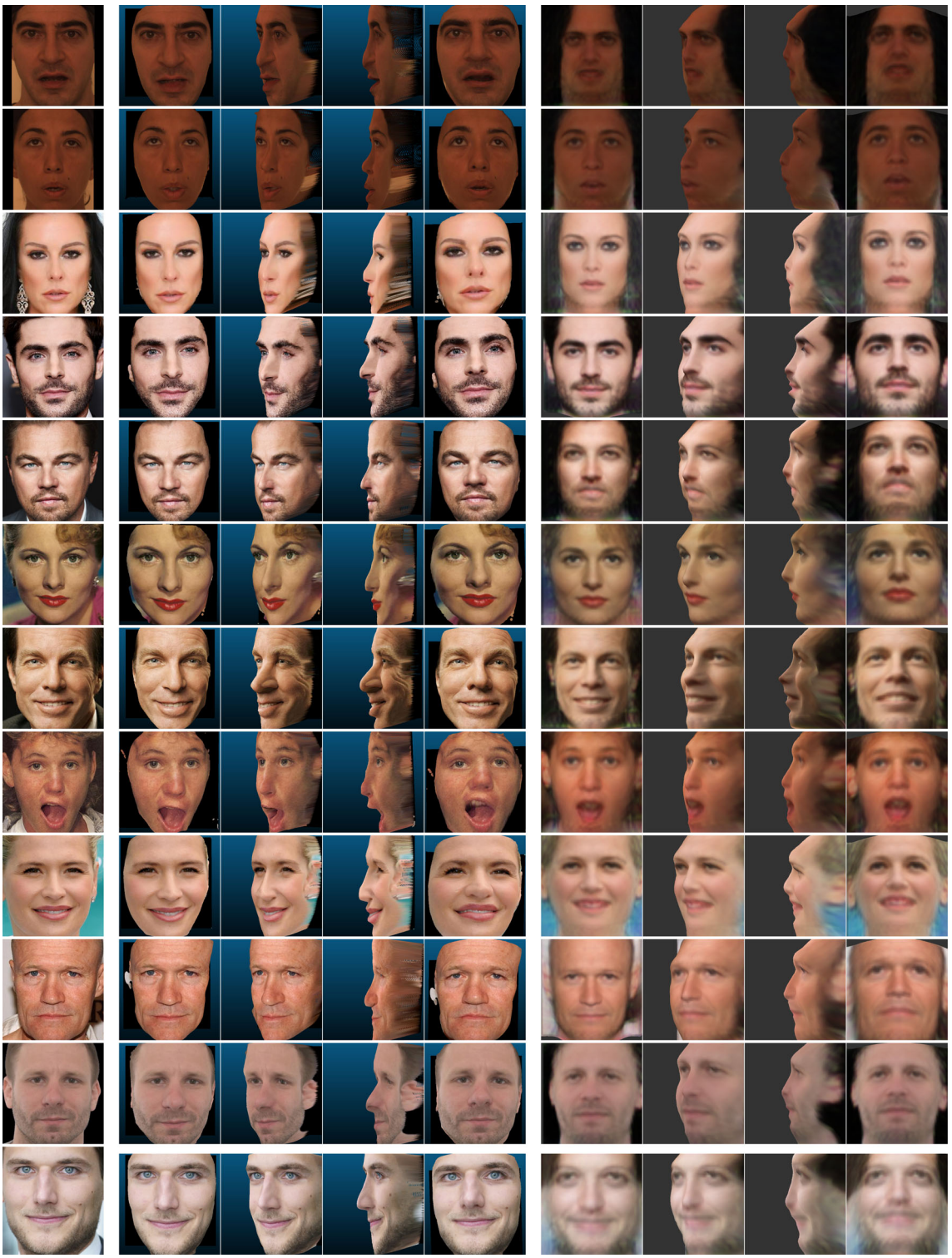

Fig. 10 From left to right: RGB input, four snapshots of the synthesized 3D model generated by our method and four snapshots of the synthesized 3D model generated by Wu et al. [26]

analysis, face recognition and segmentation algorithms. In our opinion, this database is particularly well suited for depth prediction of arbitrary portraits, as it consists of RGB images with different exposures and different image backgrounds. Furthermore, all images are already cropped to a face-bounding box. We randomly select 10k RGB frames and resize them to a dimension of 256×256. The RGB images of the Bosphorus-3DFA and all samples of the CelebAMask-HQ are used as training data for the RGB domain, and the depth profiles of the Bosphorus-3DFA are used for the depth domain. We conduct unsupervised training of our proposed framework as described in Algorithm 1. During optimization, RGB images are scaled from $[0, 255]$ to $[-1, 1]$ and depth profiles are scaled from $[0, 1]$ to $[-1, 1]$.

We qualitatively benchmark our proposed method against Wu et al. [26], where a method for fully unsupervised 3D modeling out of single images is introduced. To be more exact, a network is proposed that factors each input RGB into depth, albedo, viewpoint and illumination. In order to disentangle these different components without any supervision via paired data, stereo pairs or video sequences, Wu et al. make use of the fact that faces have in principle a symmetric structure. Thus, this proposed method for image disentanglement can also be applied to other object categories, provided that these have a symmetrical structure. The research of Wu et al. is one of the few works which has especially been developed for 3D modeling and where no supervision via paired RGB-depth data or availability of video sequences and stereo images is possible. The method has been evaluated on several databases of cat and human faces, also including the CelebA. For visual comparison, we make use of the publicly available demo version [2] provided by the authors.

We visually evaluate the success of the proposed unsupervised approach and present in Fig. 10 synthesized 3D models that were created from RGB images of the Bosphorus-3DFA, the CelebAMask-HQ, and images in the wild.

References

1. Nathan Silberman, P.K. Derek Hoiem, Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV (2012)
2. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving the KITTI vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
3. Zhao, C., Sun, Q., Zhang, C., Tang, Y., Qian, F.: Monocular depth estimation based on deep learning: an overview. Sci. China Technol. Sci **63**, 1612–1627 (2020)
4. Angermann, C., Haltmeier, M., Laubichler, C., Jónsson, S., Schwab, M., Moravová, A., Kiesling, C., Kober, M., Fimml, W.: Surface topography characterization using a simple optical device and artificial neural networks. Eng. Appl. Artif. Intell. **123**, 106337 (2023). https://doi.org/10.1016/j.engappai.2023.106337
5. Laubichler, C., Kiesling, C., Kober, M., Wimmer, A., Angermann, C., Haltmeier, M., Jónsson, S.: Quantitative cylinder liner wear assessment in large internal combustion engines using handheld optical measurement devices and deep learning. In: 18. Tagung Nachhaltigkeit in Mobilität, Transport und Energieerzeugung. IVT Mitteilungen/Reports, pp. 217–231. Verlag der Technischen Universität Graz (2021)
6. Arslan, A.T., Seke, E.: Face depth estimation with conditional generative adversarial networks. IEEE Access **7**, 23222–23231 (2019)
7. Khan, F., Basak, S., Javidnia, H., Schukat, M., Corcoran, P.: High-accuracy facial depth models derived from 3D synthetic data. In: 2020 31st Irish Signals and Systems Conference (ISSC), pp. 1–5 (2020). IEEE
8. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. In: ACM SIGGRAPH 2008 Papers, pp. 1–9 (2008)
9. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 109–117 (2017)
10. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7750–7759 (2019)
11. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3D face analysis. In: European Workshop on Biometrics and Identity Management, pp. 47–56. Springer (2008)
12. Gupta, S., Castleman, K.R., Markey, M.K., Bovik, A.C.: Texas 3D face recognition database. In: 2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI), pp. 97–100. IEEE (2010)
13. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human 3.6m: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
14. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Adv. Neural. Inf. Process. Syst. **27**, 2366–2374 (2014)
15. Kwak, D.-h, Lee, S.-h: A novel method for estimating monocular depth using cycle gan and segmentation. Sensors **20**(9), 2567 (2020)
16. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1851–1858 (2017)
17. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer vision, pp. 3828–3838 (2019)
18. Jafarian, Y., Park, H.S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12753–12762 (2021)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 27. Curran Associates, Inc., Red Hook (2014)
20. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251 (2017). https://doi.org/10.1109/ICCV.2017.244

[2] https://www.robots.ox.ac.uk/~vgg/blog/unsupervised-learning-of-probably-symmetric-deformable-3d-objects-from-images-in-the-wild.html?image=004_face&type=human.

21. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 214–223. PMLR (2017)
22. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein GANs. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., Red Hook (2017)
23. Lee, C.-H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5549–5558 (2020)
24. Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX, pp. 319–345. Springer, Berlin (2020). https://doi.org/10.1007/978-3-030-58545-7_19
25. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2422–2431 (2019). https://doi.org/10.1109/CVPR.2019.00253
26. Wu, S., Rupprecht, C., Vedaldi, A.: Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1–10 (2020)
27. Villani, C.: Optimal Transport: Old and New, vol. 338. Springer, Berlin (2008)
28. Han, X.: MR-based synthetic CT generation using a deep convolutional neural network method. Med. Phys. **44**(4), 1408–1419 (2017)
29. Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Carass, A., Prince, J.L., Sugano, N., Sato, Y.: Cross-modality image synthesis from unpaired data using cyclegan. In: International Workshop on Simulation and Synthesis in Medical Imaging, pp. 31–41. Springer (2018)
30. Lei, Y., Harms, J., Wang, T., Liu, Y., Shu, H.-K., Jani, A.B., Curran, W.J., Mao, H., Liu, T., Yang, X.: MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. Med. Phys. **46**(8), 3565–3581 (2019)
31. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2015). arXiv:1511.06434 [cs.LG]
32. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
33. Jung, H., Kim, Y., Min, D., Oh, C., Sohn, K.: Depth prediction from a single image with conditional adversarial networks. In: 2017 IEEE International Conference on Image Processing, ICIP 2017 - Proceedings, pp. 1717–1721. IEEE Computer Society (2018). https://doi.org/10.1109/ICIP.2017.8296575
34. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 239–248. IEEE (2016)
35. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2002–2011. IEEE (2018)
36. Wofk, D., Ma, F., Yang, T.J., Karaman, S., Sze, V.: FastDepth: Fast Monocular Depth Estimation on Embedded Systems. In: IEEE International Conference on Robotics and Automation (ICRA) (2019)
37. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications (2017)
38. Pilzer, A., Xu, D., Puscas, M., Ricci, E., Sebe, N.: Unsupervised adversarial depth estimation using cycled generative networks. In: 2018 International Conference on 3D Vision (3DV), pp. 587–595. IEEE (2018)
39. Zhao, S., Fu, H., Gong, M., Tao, D.: Geometry-aware symmetric domain adaptation for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9788–9798 (2019)
40. Kundu, J.N., Uppala, P.K., Pahuja, A., Babu, R.V.: Adadepth: unsupervised content congruent adaptation for depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2656–2665 (2018)
41. Zheng, C., Cham, T.-J., Cai, J.: T2net: synthetic-to-realistic translation for solving single-image depth estimation tasks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 767–783 (2018)
42. Bhoi, A.: Monocular depth estimation: a survey (2019)
43. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. Adv. Neural. Inf. Process. Syst. **29**, 658–666 (2016)
44. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
45. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Process. Syst. **30** (2017)
46. Babakhani, P., Zarei, P.: Automatic gamma correction based on average of brightness. Adv. Comput. Sci. Int. J. **4**(6), 156–159 (2015)
47. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2427–2436 (2019)
48. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
49. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283 (2016)
50. Wood, E., Baltrusaitis, T., Hewitt, C., Dziadzio, S., Cashman, T.J., Shotton, J.: Fake it till you make it: face analysis in the wild using synthetic data alone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3681–3691 (2021)
51. Yang, S., Luo, P., Loy, C.-C., Tang, X.: From facial parts responses to face detection: A deep learning approach. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3676–3684 (2015). https://doi.org/10.1109/ICCV.2015.419

Christoph Angermann received his M.Sc. in Mathematics in 2019 for his work on projection networks in biomedical volumetric segmentation. He received his Ph.D. in 2023 for his research on generative machine learning for unsupervised modality transfer and uncertainty quantification. Since then, he has been working as a postdoc at VASCage - Centre on Clinical Stroke Research.

Matthias Schwab received his master's degree in mathematics in 2021 from the University of Innsbruck. He is currently doing his PhD studies on the Medical University of Innsbruck. His main research interest is deep learning with applications to medical image analysis.

Markus Haltmeier received his Ph.D. degree in mathematics from the University of Innsbruck, Austria, in 2007, for research on computed tomography. Since 2012, he is full professor with the Department of Mathematics, University of Innsbruck. His current research interests include inverse problems, regularization theory, signal and image processing, computerized tomography, photoacoustic imaging and machine learning.

Christian Laubichler received his M.Sc degree in technical mathematic from the Technical University of Graz. From 2020, he has been working as a Data Scientist in the Digital Engine area at LEC GmbH in Graz (Austria).

Steinbjörn Jónsson received his M.Sc. in mechanical engineering in 2017 from RWTH Aachen University with focus on computer aided conception and production in mechanical engineering. From 2018, he has been working as a mechanical development engineer with INNIO Jenbacher GmbH.