Synthetic Data Generation for Demonstrating Noise Reduction in Facial Depth Imaging

Connah Kendrick
Manchester Metropolitan University
UK, Manchester, Chester Street,
Connah.Kendrick@mmu.ac.uk

Moi Hoon Yap
Manchester Metropolitan University
UK, Manchester, Chester Street,
m.yap@mmu.ac.uk

Kevin Tan
University of Inland Norway
Norway
kevin.tan@inn.no

Abstract

Deep learning has achieved remarkable success in image denoising, especially for photographic data. However, advancements in denoising depth images, particularly those captured by Time-of-Flight (ToF) sensors, have been limited due to the scarcity of clean ground truth data. This study introduces a method for generating synthetic facial depth data that closely emulates the noise characteristics of ToF sensors, facilitating the creation of paired clean and noisy datasets for supervised learning. We evaluate state-of-the-art convolutional neural networks (CNNs) on these synthetic datasets to assess their denoising performance. The findings demonstrate that synthetic datasets can effectively train depth-denoising models, thus enhancing the quality of facial depth maps in practical applications. Our results suggest that using synthetic data to create realistic, noisy, and clean datasets can highlight denoising performance through advanced techniques.

1. Introduction

Image denoising has evolved into a highly specialised domain, with convolutional neural networks (CNNs) achieving exceptional performance in mitigating noise from photographic images [21]. A pivotal factor contributing to advancements in this field is the availability of extensive datasets wherein clean images are subjected to artificial degradation with realistic noise, enabling networks to learn efficacious restoration techniques [1]. These methods have demonstrated remarkable capabilities in eliminating both real-world and synthetic noise, frequently yielding visually enhanced results.

Nevertheless, these techniques encounter increased complexity when applied to time-of-flight (ToF) sensor data, which is inherently prone to noise [24, 26]. This inherent noise complicates the acquisition of pristine images necessary for CNN training. Consequently, alternative methodologies have been devised for cleansing ToF-style data, encompassing self-supervised methods [32], temporal approaches [11, 14], and conventional image processing techniques [12]. Each approach possesses inherent limitations. For instance, temporal methods involve iterative frame combinations, potentially resulting in the blurring of rapidly moving objects. Similarly, self-supervised and traditional techniques may sacrifice intricate details in complex structures, such as facial features where subtle expressions pose significant challenges for accurate reconstruction [19]. These issues have considerably impeded progress in denoising ToF-style data relative to other imaging fields.

In this paper, we conduct a comparative analysis encompassing:

- The generation of a synthetic Kinect-based dataset of Facial Action Coded (FACs) movements utilizing realistic Kinect noise.
- The training of CNN denoising algorithms on the synthetic dataset and the evaluation of results.
- Comparison of our findings against state-of-the-art Kinect denoising algorithms.
- Demonstration of the efficacy of synthetic data generation in producing reliable ground truth datasets.

This study establishes a proof-of-concept for the utilisation of synthetic data generation to enhance the quality of facial depth maps and advocates for further investigation

into synthetic approaches within the realm of depth imaging research.

2. Related Works

The related works are divided into three domains: synthetic data generation, CNN-based denoising methodologies, and depth-data-based denoising techniques.

2.1. Synthetic Data Generation

In recent times, the generation of synthetic data has become an increasingly prevalent paradigm. Researchers have harnessed the capabilities of Generative Adversarial Networks (GANs) to generate realistic visual imagery and perform image denoising [8]. Within the domains of healthcare and robotics, simulations and synthetic data are employed to test procedures under controlled conditions prior to their application on actual patients [9,31]. Moreover, in the areas of facial animation and emotion detection, the development of 3D avatars and digital twins has become a standard practice [28].

Despite technological advances, limited research Despite technological advancements, there is a paucity of research concerning the utilisation of simulated sensors to produce synthetic datasets for real-world applications. Open-source tools such as Blensor [18] can create highly accurate virtual representations of sensor devices by replicating inherent limitations, noise characteristics [10, 26], potential interferences, and even multi-sensor noise when required. Nevertheless, these simulations encounter several challenges, such as the colour absorption properties of infrared (IR) light affecting the perception of depth and reflective surfaces. More common approaches concentrate on obtaining depth estimations from noisy sensor data. For instance, Breckon et al. [2] leveraged synthetic data and style transfer techniques to predict depth images using the KITTI dataset [15], an approach further expanded upon by Zhenyu et al. [25]. However, these methodologies remain contingent on generating synthetic data from inherently noisy sources.

Synthetic data has also been extensively investigated in facial analysis within traditional imaging. Wood et al. [36] demonstrated that synthetic data alone could facilitate robust in-the-wild facial analysis by employing a procedurally generated 3D facial model that simulates identity, expression, clothing, hair, lighting, and environmental backgrounds. Their research focused on two crucial aspects: face parsing (segmentation of key facial regions) and landmark localisation (detection of essential facial features). Additionally, Jiang et al. [4] utilised synthetic data for dataset balancing in kinship recognition, combining facial alignment, feature extraction, and style transfer to address imbalance. More recently, George et al. [16] introduced a pipeline to enhance synthetic data generation for facial recognition. Their method refined existing synthetic images by generating prototypes with class variations and integrating video clips to boost image realism and recognition accuracy. However, the relatively limited focus on the application of synthetic depth data remains a notable deficiency in the current literature.

2.2. denoising

Depth data is occasionally utilised in conjunction with real-world noisy sensor datasets within this domain. For instance, Tong et al. [34] have presented an image compression technique that incorporates depth information by analysing light wave reflections, facilitating reflectance calculations. Notwithstanding, contemporary methodologies predominantly depend on synthetic data or alternative capture techniques. This section evaluates these datasets to underscore the necessity for an established gold standard of depth data to support ToF denoising.

The DIV2K dataset [1], proposed by Eirukur et al., remains the most prevalently employed dataset. This dataset emphasizes image super-resolution and downsampling, alongside denoising. The authors employ bi-cubic downsampling and introduce unspecified noise augmentations to deliberately degrade image quality. DIV2K amalgamates several conventional benchmark denoising datasets—namely Set5 [33], Train91 [22], Set14 [40], B100 [27], and Urban100 [20]—thereby facilitating model generalization and satisfying the extensive data requirements of deep learning.

In the current context, the most advanced approach is a self-supervised method based on Gaussian noise, as proposed by Monroy et al. [29]. Additionally, the High-Quality Denoising Dataset for Smartphone Cameras furnishes images captured from five smartphones under varying lighting conditions. Multiple frames are combined and corrected to generate a semi-synthetic ground truth for the final image. The cascade gaze-style CNN proposed by Ghasemabadi et al. [17] demonstrates superior performance on this dataset.

Another significant dataset is the See-in-the-Dark dataset [6], which provides two images per scene, a short-exposure and a long-exposure capture, to facilitate the mitigation of noise associated with high ISO settings and low-light conditions. Here, the most advanced technique employs a physics-based approach that explicitly incorporates the physical attributes contributing to noise in low-light photography [21].

It is noteworthy that numerous works employing alternative datasets artificially introduce noise using standard image processing techniques, such as downsampling followed by resizing, the addition of salt-and-pepper noise, and the application of various Gaussian blurs to simulate synthetic noise [21], an option not available for standard ToF data.

2.3. Depth Data denoising

Depth data denoising frequently employs temporal methods such as those detailed by Luo et al. [23]. Their approach utilizes longitudinal depth maps and alignment via ICP algorithms, potentially incorporating accelerometer data to enhance accuracy. The depth frames are integrated by averaging overlapping regions, thereby mitigating noise. In both commercial and high-end packages, this methodology is extended through the use of 3D Morphable models and deformable models, effectively yielding cleaner models.

Given the limited availability of ground truth data, self-supervised methods have been widely adopted. Sterzentsenko et al. [32] developed a fully convolutional neural network (CNN) that enhances depth maps beyond the intrinsic limitations of sensor data. Their system, tested with ToF sensors such as the Kinect V2 and Realsense D415, focuses on scene-wide denoising by merging multiple scene views to create a singular, clean map.

In contrast, Mu et al. [30] employed a supervised method for denoising depth data. They utilised high-accuracy scanner datasets of faces, introducing noise through Gaussian distribution and simple downsampling/upsampling techniques to generate low-quality models. Their process includes facial cropping via nose point detection, hole filling, and outlier removal based on neighbouring data. Though primarily focused on facial recognition, their study does address some denoising steps. However, the artificially generated noise does not accurately represent ToF sensor noise. Xu et al. [37] expanded upon this work, enhancing normal map generation and developing a multi-modality network, while employing a similar dataset generation strategy. Yet, their comparison with traditional denoising techniques remains limited.

Recent studies have also explored RGB-driven denoising, wherein colour signals are leveraged to smooth depth maps, as demonstrated by Sterzentsenko et al. [32] and Yan et al. [38]. These techniques map standard colour photography onto depth images, using image smoothness to reduce depth map noise.

2.4. Evaluation

We use their metrics to evaluate our models; for loss, we implement MSE 1. Additionally, we provide PSNR 2 and SSIM 3.

$$MSE = \sum_{i=0}^{n} \frac{(y_i - y_i')^2}{n} \tag{1}$$

where:

- n is the number of samples in the training batches.
- y_i is the ground truth for the training image.
- y_i' is the predicted output for the training image.

$$\text{PSNR} = 20 \times \log_{10}(\text{MAX}) - 10 \times \log_{10}(\text{MSE}) \tag{2}$$

where:

- MAX is the maximum possible value in the ground truth; for ours it is 8000.
- MSE is equation 1.

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\alpha_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\alpha_x^2 + \alpha_y^2 + c_)} \tag{3}$$

where:

- μ_x is the average of the input image x
- μ_y is the average of the input image y
- α_x^2 is the variance of the input image x
- α_y^2 is the variance of the input image y
- α_{xy} is the covariance of the input images x and y
- $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ are used to stabilise the division with a weaker denominator where:
  - L is the dynamic range found in the pixels
  - k_1 = 0.001, k_2 = 0.003 by default

3. Methodology

For our technique, we generated a synthetic Kinect-based dataset from existing 3D facial datasets, namely:

- *D3DFACs [7]:* Is a FACs coded dataset of 3D models performing Action Units (AUs) from onset-peak-offset. The dataset contains 10 subjects displaying between 19 to 97 different action units, creating over 519 total AUs. The participants were recorded using a 3DMD dynamic 3D stereo camera, allowing the capture of the full expressions in high accuracy at 60 FPS.
- *Face Warehouse [5]:* Generate a 3D face dataset of 150 participants recorded with the Kinect. The participants created clean models posing a series of 19 facial expressions and using temporal integration with RGB data. Recent works have improved the technique to be single frame [23]
- *Biwi Kinect [13]:* Is recorded for 20 people to access head movement tracking. The dataset was produced directly due to the difficulty in properly cleaning depth depth, causing head pose estimation difficulties.

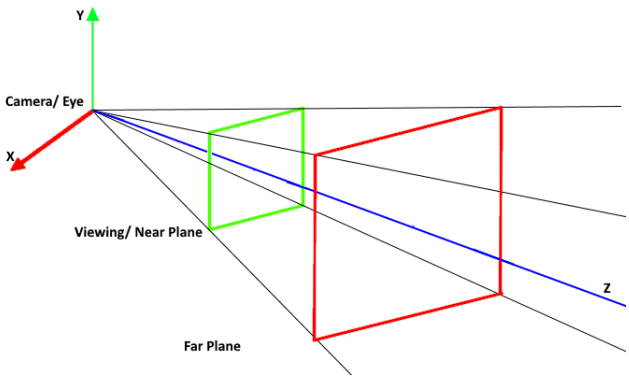

Figure 1. The viewing frustum use for ray tracing the models

Using our generated ground truth images, we employ this synthetic dataset to evaluate the performance of standard depth data denoising algorithms in comparison with traditional denoising datasets. This approach exemplifies an active application wherein large-scale synthetic data is produced that mirrors real-world challenges.

3.1. Training

3.2. Dataset Generation

We ensure a rigorous dataset split by participants to prevent overlap between training, validation, and test sets, thus showcasing the model's ability to generalize. Subsequently, we processed these images through a modified version of Blensor [18] and render-Kinect [3], specifically targeting the noise characteristics of the Kinect V2 sensor to achieve realistic noise generation. To simulate this, we employ Ray-Tracing from the sensor origin into world space, ensuring the sensor remains within the simulated viewing frustum 1, and checking for intersection with the models in the environment. The ray-trace algorithm operates by tracking an intersection between an Origin point with a direction 4 and a Polygon surface 5, providing the relative vector from the origin.

$$P(t) = O + tD, T \geq 0 \tag{4}$$

Where:

- O is the sensor origin
- D is the direction and $T \geq 0$ is the distance

$$t = \frac{(N.V) - (N.O)}{N.O} \tag{5}$$

Where:

- N is the surface normal
- V is the surface vertex

The system accurately considers sensor occlusion from the IR Projector and IR camera, resulting in depth map shadowing as described by Mallick et al. (2014) using equation 6. Additionally, the system incorporates sensor image quantization and utilizes Gaussian simplex and Perlin noise functionalities to realistically model the ray-tracing capabilities of Kinect data. This comprehensive approach ensures a faithful representation of various Kinect noise models, including axial, shadow, and lateral noise, albeit excluding surface noise. We have modified the system to simultaneously output both clean and noisy depth maps. For each model loaded into the system, we generate two 16-bit UShort images: a clean image mapping to the Kinect's 512×424 depth resolution and a noisy image incorporating the processing steps that simulate Kinect noise. During testing, each facial model was positioned to face the virtual sensor, centrally aligned within the field of view at a distance of approximately 90 cm, with no additional background noise introduced.

$$b = bf(1/Z_o - 1/Z_b) \tag{6}$$

where:

- d is the distance from the sensor.
- bf is the image plane and sensor depth.
- Z_o and Z_b are the distance between the image plane of the object and the background, respectively.

We synthetically generated $32,029$ training images, $5,503$ validation images, and $8,406$ test images; the split was based on participants to ensure no overlap occurred between training, validation, or test sets. To prepare the data for network use, we performed a centre crop of the depth mask to focus on facial regions at a resolution of 128×128 pixels. To preserve spatial consistency, we did not resize the images. We automatically positioned the faces at 1 millimetre in image space to facilitate model training.

To compare techniques, we replicated the state-of-the-art methods for the DIV2K dataset described by Monroy et al. [29]. We used the Adam optimiser with a learning rate of 1×10^{-4}, a batch size of 4, and early stopping with a patience parameter of 10 epochs. Additionally, we extended their experiments by incorporating more commonly used versions of the DnCNN model with 20 and 64 depth layers.

We did not reproduce the results of the "See in the Dark" approach, as the noise generated by ToF sensors differs significantly from that produced by CMOS camera sensors. Furthermore, our focus on facial data precludes the use of temporal alignment methods, which may obscure facial expressions at low frame rates. Similarly, we avoided networks requiring camera parameters or self-supervised tech-

Table 1. This table highlights that the most commonly shown metrics in depth de-noising are prone to showing a near perfect result, whereas the lesser shown metrics highlight significant issues.

Filters	Median	NLM	Adaptive	Bilateral	Baseline
SSIM	0.9986	0.8373	0.9986	0.7704	0.9986
PSNR	47.7992	48.4857	47.6132	47.8734	47.6132
MSE	1361.1435	1147.3265	1410.8397	1349.4505	1410.8397

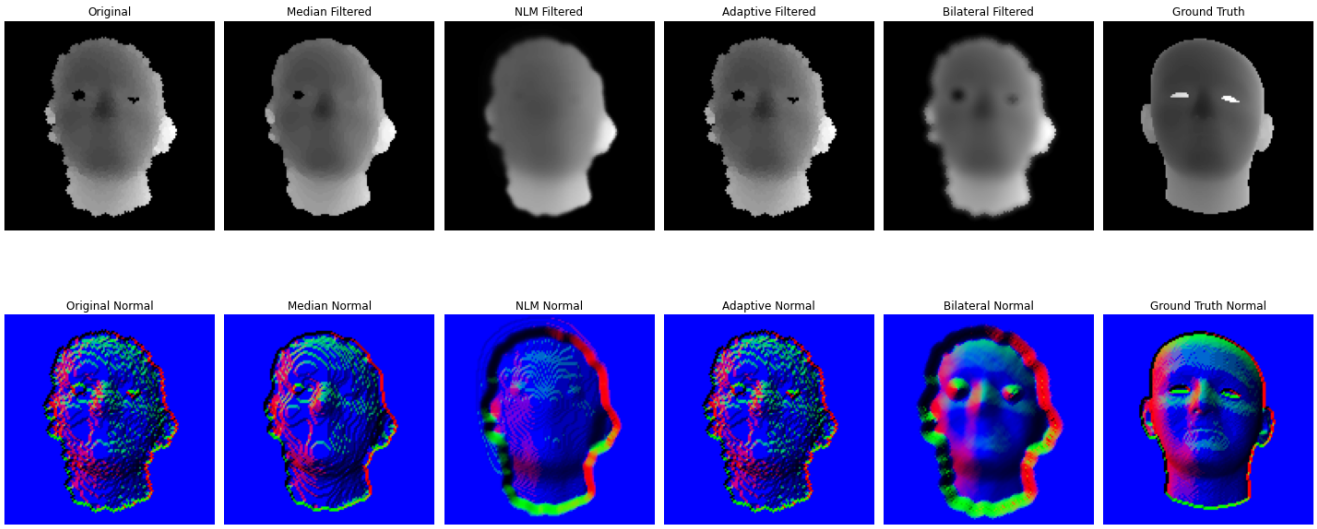

Figure 2. A comparison of different traditional methods on depth maps, compared to a ground-truth image.

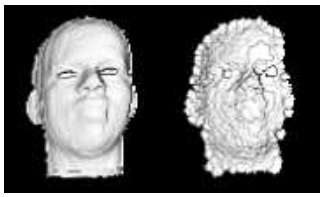

Figure 3. An example rendering of the FaceWarehouse model depth maps rendered in 3D with the left our clean, and right the synthetic noise simulating Kinect data.

niques, instead demonstrating the practical value of the synthetic dataset with the current state-of-the-art image denoising methods.

However, we incorporated traditional techniques described by Essmael et al. [12]. Specifically:

- For the Median filter, we used a kernel size of 5.
- For the Bilateral filter, we employed a diameter of 9, a sigma colour of 75, and a sigma space of 75.
- For NLM, we specified a filter strength of 30, a window size of 7, and a search window size of 21.
- For adaptive thresholding, we set the threshold value to 15.

Finally, we benchmarked the results using several state-of-the-art image denoising techniques to underscore the advantages of the Synthetic Generation approach.

4. Results

For traditional techniques, we observe that certain metrics can present an obscured view of system performance. As highlighted in Table 1, high SSIM scores may suggest that traditional approaches provide high-quality denoising resembling the ground truth. However, other metrics reveal significant discrepancies in quality. This highlights a fundamental issue with traditional metrics for denoising depth data, where L is computed to ensure that low variance of α_x^2 and α_y^2 is less weighted in standard images. Consequently, the millimetre range of $0-8000$ leads to near-perfect scores. Whereas, with MSE, by squaring the values, even minor millimetre differences transparently indicate denoising errors.

Table 1 also shows that traditional techniques struggle due to quality improvements that affect the baseline column, contrasting noisy and clean images. This minimal impact is visually reflected in Fig. 5, where traditional techniques show limited effectiveness when compared against the availability of ground truth data. Similarly, we focus on typical regions of facial denoising and demonstrate that

Table 2. The results of Transformer and Unet style networks, demonstrating low performance.

Models	Restormer	UNet
SSIM	0.9733	0.9673
PSNR	29.1483	30.5461
MSE	80017.7539	59091.5583

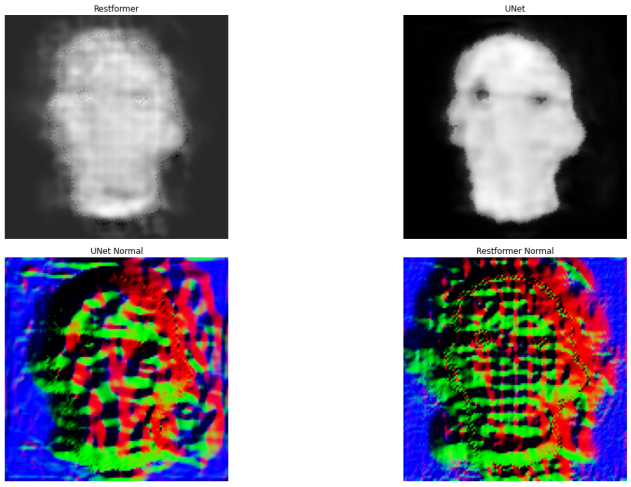

Figure 4. A comparison of different transformer-based architecture and UNet on depth maps.

widely used traditional methods fail to produce realistic results. In the bottom row of Fig. 5, we generate normal maps of the results, illustrating how the Bilateral filter creates a smooth surface with anomalies around the eyes, while other techniques retain more jagged edges but are visually improved over deep learning techniques.

The differences in scores between our method and those reported by Essmaeel et al. [12] can be attributed to their evaluation on flat surfaces, which contrasts sharply with the complexity of facial features.

We compare our model against the state-of-the-art approach described in Monroy et al. [29], under a range of conditions detailed in Table 3. Here, D represents the depth of the network, and AR refers to the inclusion of Artefact Removal. Through metric analysis, traditional methods using SSIM and PSNR demonstrate seemingly good performance, in stark contrast to negative results observed with MSE. Additionally, we find that CNNs struggle significantly with this type of data; while overall performance remains low, artefact removal aids shallow networks.

This phenomenon is further amplified in Fig. 5, where the performance improvement does not correlate with visual enhancements. This discrepancy arises from the nature of depth data. In this case, out-of-bound values are set to 0, which causes convolutional features to form a concave effect as higher values are pulled downward by zeros. This concavity is further detailed in deeper networks and is emphasised in artefact removal processes.

We expand our comparison to include models less susceptible to convolutional bluring, such as Visual Transformers and U-Net-based structures, Restormer [39] and a Unet structure in which early network outputs are combined with later. However, as shown in Table 2, results follow the same trajectory as other analyses. Likewise a visual comparison **??** highlights that the use of deep learning for the analysis of depth data is still far from usable for high end applications.

Owing to the generation of synthetic data, we can see that standard image denoising techniques do not transfer into the field of depth data. In addition, due to the background constraints, smoother results tend to create an uneven board around the facial region. Generating synthetic data can help us improve and develop suitable denoising methods; it highlights that currently, traditional techniques are more beneficial than the Deep-CNN architectures and remains an open avenue of future work to allow for increased performance. Experiments reflected that some methods created invalid depth images with a positive metric performance. This highlights the issues of MSE and follows previous studies with the metric. [35].

5. Conclusion

Our experiments demonstrated that CNN-based methods trained on this synthetic data could not outperform traditional filtering techniques, especially in preserving fine facial structures, due to the nature of the convolutional networks. This highlights the need for Synthetic data to create benchmarks for future depth-data analysis. Furthermore, we showed that synthetic datasets can serve as effective stand-ins for real-world ground truth, addressing the limitations caused by the lack of clean-depth data.

6. Future works

By implementing this methodology, techniques demonstrated by Woods et al. [36], to include whole body and background, could significantly improve the field but require a significant amount of computational power.

Acknowledgement

We want to thank Cao et al. [5], Fanelli et al. [13], and Cosker et al. [7] for contributing to open research by providing their datasets, which would not have made this work possible.

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In

Table 3. The results of the DnCNN models, where D indicates the model depth and AR indicates an artifact removal.

Models	DnCNN D7	DnCNN D20	DnCNN D64	DnCNN AR D7	DnCNN AR D20	DnCNN AR D64
SSIM	0.9861	0.9852	0.9885	0.9838	0.9838	0.9881
PSNR	32.9725	33.2791	33.3734	33.3941	33.2898	33.2267
MSE	36238.2963	34237.0065	33905.1296	33701.7603	34235.7125	34459.7883

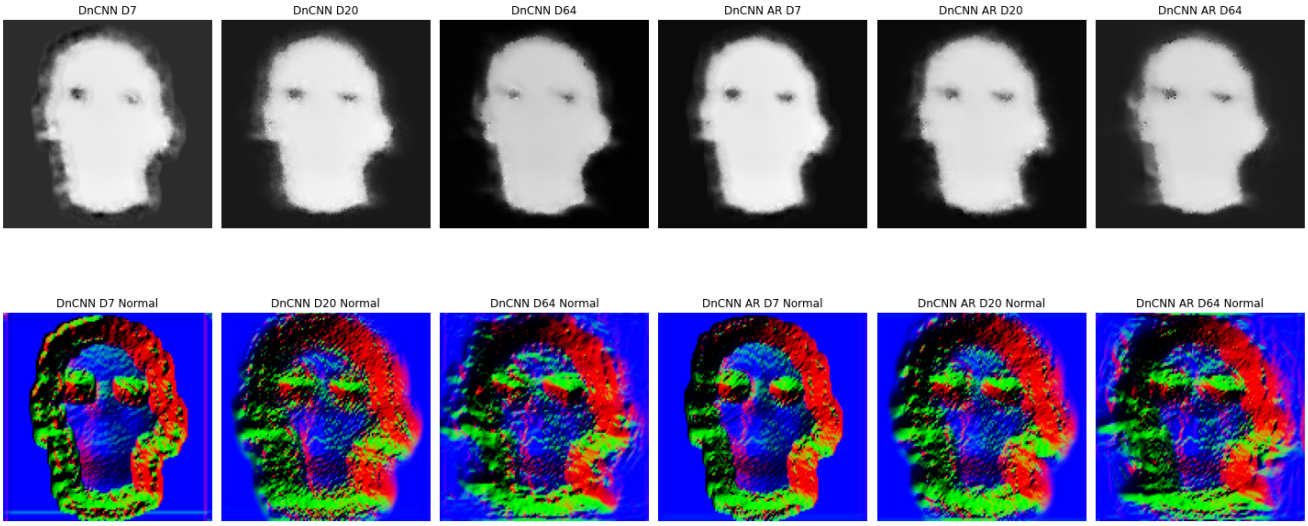

Figure 5. A comparison of different DnCNN methods on depth maps.

2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1122–1131, 2017. 1, 2

[2] A. Atapour-Abarghouei and T. P. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2800–2810, 2018. 2

[3] J. Bohg, J. Romero, A. Herzog, and S. Schaal. Robot arm pose estimation through pixel-wise part classification. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3143–3150. IEEE, 2014. 4

[4] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, 135:104688, 2023. 2

[5] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 3, 6

[6] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 2

[7] D. Cosker, E. Krumhuber, and A. Hilton. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2296–2303, 2011. 3, 6

[8] R. Dabral, M. H. Mughal, V. Golyanik, and C. Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9760–9770, 2023. 2

[9] R. Daher, C. Kendrick, M. H. Yap, D. Leff, and S. Giannarou. Vision-based robot localisation for ductoscopic navigation. In *Hamlyn Symposium on Medical Robotics*, page 89. 2

[10] N. M. Difilippo, M. K. Jouaneh, and S. Member. Characterization of Different Microsoft Kinect Sensor Models. 15(8):4554–4564, 2015. 2

[11] K. Essmaeel, L. Gallo, E. Damiani, G. De Pietro, and A. Dipand?? Temporal denoising of Kinect depth data. *8th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2012r*, pages 47–52, 2012. 1

[12] K. Essmaeel, L. Gallo, E. Damiani, G. De Pietro, and A. Dipanda. Comparative evaluation of methods for filtering kinect depth data. *Multimedia Tools and Applications*, 74:7331–7354, 2015. 1, 5, 6

[13] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random Forests for Real Time 3D Face Analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013. 3, 6

[14] J. Fu, S. Wang, Y. Lu, S. Li, and W. Zeng. Kinect-like depth denoising. In *2012 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 512–515, 2012. 1

[15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 2

[16] A. George and S. Marcel. Digi2real: Bridging the realism gap in synthetic data face recognition via foundation models. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1469–1478, 2025. 2

[17] A. Ghasemabadi, M. K. Janjua, M. Salameh, C. Zhou, F. Sun, and D. Niu. Cascadedgaze: Efficiency in global context extraction for image restoration. *arXiv preprint arXiv:2401.15235*, 2024. 2

[18] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree. BlenSor: Blender Sensor Simulation Toolbox. 6939(Isvc):199–208, 2011. 2, 4

[19] Y. Guo, F. Luo, and S. Xu. Self-supervised face image restoration with a one-shot reference. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2930–2934, 2024. 1

[20] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 2

[21] A. E. Ilesanmi and T. O. Ilesanmi. Methods for image denoising using convolutional neural network: a review. *Complex & Intelligent Systems*, 7(5):2179–2198, 2021. 1, 2

[22] Y. Jianchao, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2

[23] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu. 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, 27(10):4756–4770, 2018. 3

[24] C. Kendrick, K. Tan, T. Williams, and M. H. Yap. An Online Tool for the Annotation of 3D Models. pages 362–369, 2017. 1

[25] Z. Li, X. Wang, X. Liu, and J. Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *IEEE Transactions on Image Processing*, 33:3964–3976, 2024. 2

[26] T. Mallick, P. P. Das, and A. K. Majumdar. Characterizations of noise in kinect depth images: A review. *IEEE Sensors Journal*, 14(6):1731–1740, 2014. 1, 2

[27] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 2

[28] "Microsoft". Microsoft Kinect, 2013. 2

[29] B. Monroy, J. Bacca, and J. Tachella. Generalized recorrupted-to-recorrupted: Self-supervised learning beyond gaussian noise. *arXiv preprint arXiv:2412.04648*, 2024. 2, 4, 6

[30] G. Mu, D. Huang, G. Hu, J. Sun, and Y. Wang. Led3d: A lightweight and efficient deep approach to recognizing low-quality 3d faces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5773–5782, 2019. 3

[31] K. Singh, T. Navaratnam, J. Holmer, S. Schaub-Meyer, and S. Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2505–2515, 2024. 2

[32] V. Sterzentsenko, L. Saroglou, A. Chatzitofis, S. Thermos, N. Zioulis, A. Doumanoglou, D. Zarpalas, and P. Daras. Self-supervised deep depth denoising. In *ICCV*, 2019. 1, 3

[33] R. Timofte, V. De Smet, and L. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927, 2013. 2

[34] K. Tong, X. Jin, C. Wang, and F. Jiang. Sadn: Learned light field image compression with spatial-angular decorrelation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP), pages 1870–1874, 2022. 2

[35] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009. 6

[36] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton. 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*, pages 3681–3691, 2021. 2, 6

[37] R. Xu, K. Wang, C. Deng, M. Wang, X. Chen, W. Huang, J. Feng, and W. Deng. Depth map denoising network and lightweight fusion network for enhanced 3d face recognition. *Pattern Recognition*, 145:109936, 2024. 3

[38] S. Yan, C. Wu, L. Wang, F. Xu, L. An, K. Guo, and Y. Liu. Ddrnet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In *Proceedings of the European conference on computer vision (ECCV)*, pages 151–167, 2018. 3

[39] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 6

[40] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 2