Full | DOI:1234.5678 | Vol.1 No.1 (2025)

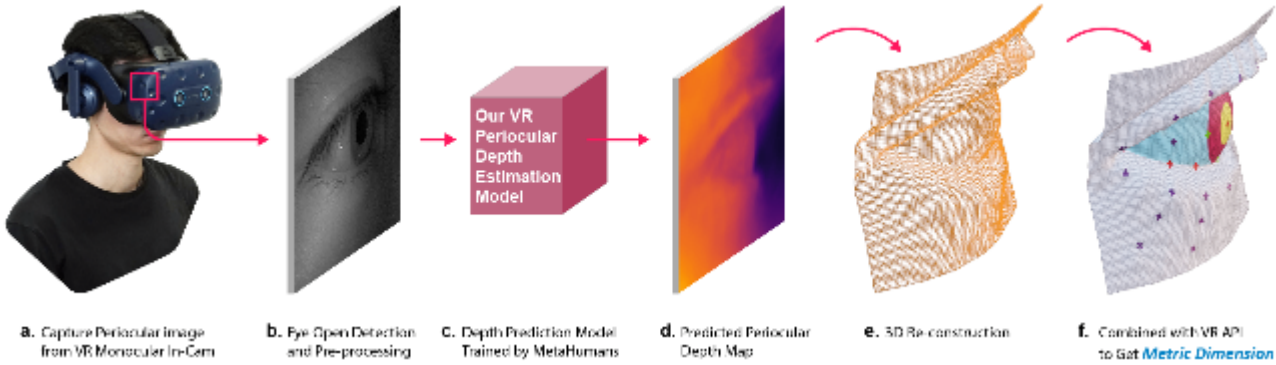
DeepMetricEye - Metric Depth Estimation in Periocular VR Imagery

Yitong Sun
yitong.sun@rca.ac.uk
Royal College of Art

Zijian Zhou
z.zhou-89@sms.ed.ac.uk
University of Edinburgh

Cyriel Diels
cyriel.diels@rca.ac.uk
Royal College of Art

Ali Asadipour*
ali.asadipour@rca.ac.uk
Royal College of Art



Abstract: VR headsets often cause eye strain, yet eye-facing cameras only supply relative depth. We introduce a lightweight U-Net 3+ model that predicts metric 3-D periocular depth from a single monocular camera on any headset, enabling precise light-stimulus estimation. To sidestep scarce real data, we build a Dynamic Periocular Data Generation pipeline with UE MetaHuman, producing thousands of synthetic images from limited scans. Tested on 36 participants, the system delivers high global depth accuracy and trustworthy pupil-diameter readings, offering a practical basis for VR eye-health assessment.

Tags: Computer vision, HCI

1 Introduction

Virtual Reality (VR) technology has advanced rapidly, offering immersive experiences across applications including gaming, medical, education, and training simulations*(Dincelli & Yayla, 2022; Zhang et al., 2022)*. The recent re-imagination of the 'Digital Universe' concept has illuminated a compelling vision of globally interconnected interactions *(Mystakidis, 2022; Ning et al., 2023)*. Despite the continual enhancement in content quality, the use of VR headsets persists in causing physiological discomfort to users, leading to substantially reduced usage duration *(Hirzle et al., 2022; Keshavarz & Golding, 2022)*. A significant proportion of unsatisfactory experiences by VR users can be attributed to digital eye strain (DES), dry eye, and visual impairment resulting from excessive artificial light stimulation from VR displays, as well as periocular swelling, increased intraocular pressure, and muscle displacement induced by the pressure exerted by the headset's face mask *(Souchet et al., 2023)*. However, these visual health issues have not yet received attention proportionate to the development of VR technology.

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

Recent VR headsets, increasingly equipped with eye-oriented monocular cameras, are designed to segment periocular feature maps, annotate the edge of the pupil, and detect gaze direction to enhance content interaction *(Fuhl et al., 2021; Li et al., 2021)*. While these methods offer a preliminary insight into eye activity during VR usage, they are insufficient for establishing connections with medical standards, for instance, the light stimulus calculation protocols and periocular condition medical guidelines, needed for meticulous visual health assessments and advanced user interaction studies *(Feder et al., 2016; on Illumination, 2018)*. The fundamental issue lies in the inability of current methods to convert the segmented 2D relative feature annotations (such as pupil edge segmentation) into spatial metrics (pupil diameter), essential for strict standards. Proposed solutions, for instance, incorporating stereo cameras and depth cameras for metric size acquisition, present substantial challenges in terms of cost, computational power, battery life, and hardware design of VR headsets.

To convert 2D periocular feature annotations into 3D metric dimensions, we propose a framework that only utilises an eye-oriented monocular camera, present in various VR headsets, to estimate the measurable periocular depth map. This framework, built on a U-Net 3+ deep learning backbone, re-optimised by us, aims to accurately estimate depth maps while maintaining the lightweight processing demands suitable for VR deployment *(Huang et al., 2020)*. To alleviate the difficulty in collecting facial data for training, we introduce a Dynamic Periocular Data Generation (DPDG) environment that leverages a small quantity of real facial scan data to generate thousands of synthetic periocular images and corresponding ground truth depth maps using Unreal Engine (UE) MetaHuman *(Fang et al., 2021)*. Figure 1 provides the snapshot of this paper.

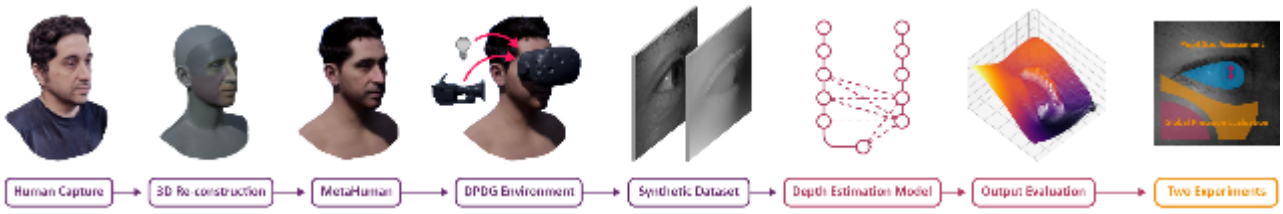


Figure 1: A Snapshot of the paper.

The main contributions of this study are as follows :

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

- We introduce a lightweight depth estimation framework for VR headsets to reconstruct periocular depth maps. The aim is to provide features' metric size for light stimulus standards calculation and periocular condition monitoring.
- Addressing the challenge of facial data collection, our DPDG environment, based on UE MetaHuman, generates thousands of periocular training images and depth maps from limited facial scans.
- We evaluate our method's accuracy and usability with two tasks: 1) evaluating global precision of periocular area , and 2) assessing pupil diameter.
- We have open-sourced the DPDG environment, the code and dataset for the depth estimation model, and all metadata from the experiments.

2 Related Work

2.1 Depth Estimation

Depth estimation is vital in computer vision, with broad applications such as autonomous driving and VR indoor space reconstruction *(Zhao et al., 2020)*. Traditional methods rely on stereo vision, estimating depth from disparity across multiple camera views, albeit with high computational demands and need for accurate camera calibration *(Laga et al., 2020)*. Active depth estimation technologies, using lasers or structured light to generate depth point clouds, offer high accuracy but suffer from slow scanning speeds, high computational needs, and large structure volumes *(Cheng et al., 2018; S. Zhang, 2018)*.

The emergence of image-based depth estimation has sparked research interest *(Ming et al., 2021)*. Eigen et al. pioneered the deep learning approach for monocular image-based depth estimation using an encoder-decoder network to transform RGB images into depth maps *(Eigen et al., 2014)*. Subsequent advances include the integration of features like transformers, attention, and residual connections, enhancing model performance but increasing computational demands *(Chen et al., 2019; Jin et al., 2021; Z. Li et al., 2021; Xu et al., 2018)*.

U-Net, initially designed for medical image segmentation, found application in depth estimation due to its simplicity and practicality *(Ronneberger et al., 2015; Siddique et al., 2021; Zhou & Kaess, 2020)*. U-Net++, proposed by Zongwei et al., introduced dense skip connections but fell short of fully exploiting multiscale information. To address this, U-Net 3+ was proposed by Huimin et al., using dense skip connections across all encoder and decoder layers, reducing parameter count while improving accuracy *(Huang et al., 2020)*.

For lightweight, robust depth estimation suitable for VR deployment, we chose U-Net 3+ as our backbone, removing shallow dense skip connections to focus on skin and eye surfaces. Deep decoder layers were weighted differently to control detail decoding, aiming to reduce noise details

such as eyebrows and eyelashes.

2.2 Eye and Periocular Feature Segmentation

Eye tracking has widespread applications in sectors such as advertising, healthcare, education, and gaming *(Holmqvist et al., 2023)*. Within VR, gaze detection has been pivotal in identifying users' areas of interest and enhancing content interaction *(Modi & Singh, 2021)*. Furthermore, it's extensively utilised in research and healthcare for understanding attention, behavioural patterns, and emotional shifts *(Kang et al., 2020; Lim et al., 2020; Lisk et al., 2020)*. Periocular feature segmentation has also been researched to reconstruct facial expressions, acquire iris information, and locate pupils *(Fuhl et al., 2019, 2022; Wang & Ji, 2017)*.

Deep learning has substantially improved periocular feature segmentation accuracy and efficiency compared to traditional methods *(Chaudhary et al., 2019; Min-Allah et al., 2021; Yiu et al., 2019)*. For instance, the EyeNet framework, with residual units and attention blocks, significantly reduced recognition rate drops caused by flickering and blur noise *(Kansal & Devanathan, 2019)*. GazeNet provides end-to-end eye movement detection, focusing on automatic event annotation without pre-labeling *(Zemblys et al., 2019)*.

Despite these advancements, most methods focus on two-dimensional image features, which can't be translated into metrical dimensions in three-dimensional space. This limits the monitoring of eye conditions according to optical standards and medical guidelines, especially given the visual fatigue and impairment caused by VR headsets.

Our method estimates a metrically measurable depth map of the periocular region using monocular images, then combines it with eye segmentation techniques to yield arbitrary feature measurements. This approach allows for effective quantification and monitoring of stimulations on the human eye from VR screen light intrusion and changes caused by the mask pressure on the periocular region.

2.3 Standards and Guidelines for Periocular Conditions

Adverse effects on the human eye from VR screens, such as DES, dryness, and visual impairment due to excessive light stimulation, are becoming a focal point of research *(Hirzle et al., 2022; Souchet et al., 2022)*. The face mask's pressure can cause elevated intraocular pressure, nerve pressure, and periorbital swelling *(Mehrfard et al., 2019)*. However, despite various studies, no unified method or standard exists for evaluating VR headsets' optical stimuli *(Hirzle et al., 2022; Y. Wang et al., 2019; Wolffsohn et al., 2023)*. This gap largely stems from the need for accurate pupil area measurements to calculate photoreceptor activation according to International Commission on Illumination (CIE) and American National Standards Institute (ANSI) standards, which current VR headsets' built-in cameras cannot perform *(on Illumination, 2018, 2022; RP-3-13, 2014)*. Furthermore, the high contrast in VR headsets complicates stimuli assessment.

Concerning periocular pressure, preliminary studies confirm the negative impacts of varying pressure levels on ocular physiology and immersion *(Y. Chen et al., 2021)*. The comprehensive Adult Medical Eye Evaluation Preferred Practice Pattern (PPP) guidelines outline measurement methods related to periocular symptoms, including muscle displacement and swelling associated with VR usage *(Feder et al., 2016)*. Yet, the inability of VR's eye-oriented cameras to convert 2D images into the 3D feature information required by PPP poses a significant challenge.

Our method mitigates this by estimating the periocular area's depth map, converting 2D images into measurable 3D facial reconstructions, and providing spatial metric information aligned with various assessment standards.

3 Method

3.1 Dataset Creation

Objective and Dataset Challenge The aim of our approach is to predict the depth map based on periocular images captured by the monocular camera within VR headsets, thereby transforming relative facial feature dimensions into metric dimensions. The supervised deep learning training requires two categories of images: 1) periocular images captured by the VR headset's internal camera, and 2) corresponding depth maps used as ground truth for training. However, our analysis of existing periocular image databases revealed a discrepancy as none met the prerequisites for training our model *(Zanlorensi et al., 2022)*. Primarily, databases utilised for gaze, pupil and iris edge detection training do not contain associated depth maps. Moreover, their data were not obtained from the internal camera of VR headsets, which prevents effective calibration of the field of view (FOV), lens distortion, and noise distribution. Additionally, given the compact space within VR headsets, deploying a depth camera to capture depth maps that align with the position of the real camera poses a considerable challenge.

Dataset Generation In response to these challenges, we created a Dynamic Periocular Data Generation (DPDG) environment using UE's MetaHuman Figure 2. Using this approach, we synthesised virtual humans from real scans, significant mitigating data collection challenges and enhancing dataset diversity. As a state-of-the-art virtual human system, MetaHuman facilitates precise facial feature simulations and can emulate dynamic eye movements, pupil dilation, and blinking. Initially, we procured 68 official MetaHuman avatars and subsequently scanned 52 real humans using RealityCapture to create additional MetaHumans by fitting facial meshes in UE 5 Figure 3 a-c) *(Reality, 2023)*. To enhance the generalisability of the dataset, these 120 MetaHumans underwent a feature-mixing process using MetaHuman Creator, culminating in the generation of 1150 unique avatars *(Fang et al., 2021)*. Using DPDG's feature blending, we combined diverse makeup levels with various skin tones for realism, bolstering dataset robustness.

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

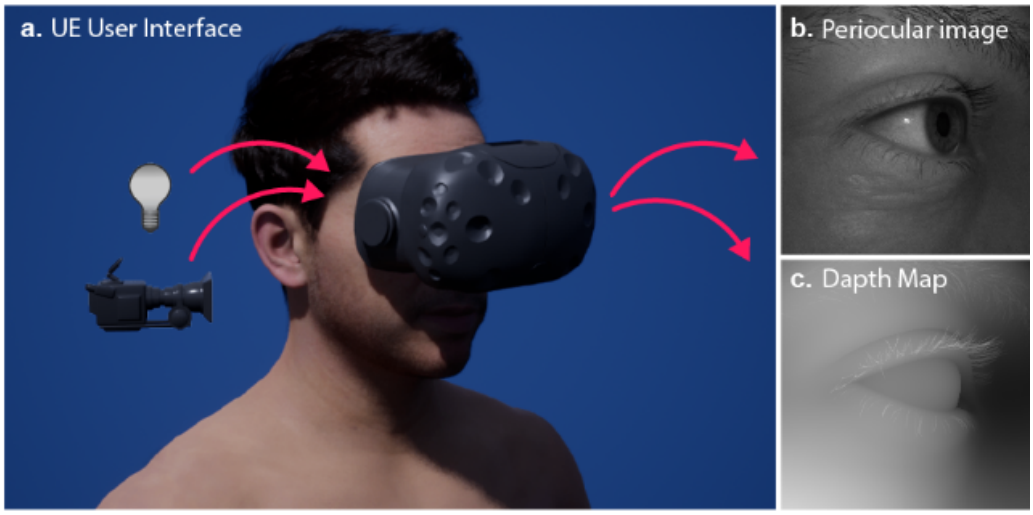


Figure 2: The Dynamic Periocular Data Generation (DPDG) environment. **a:** Metahumans are worn a specified model of VR headset with faux realistic lighting and camera in UE for periocular image acquisition. **b:** A captured synthetic periocular image. **c:**The corresponding depth map.

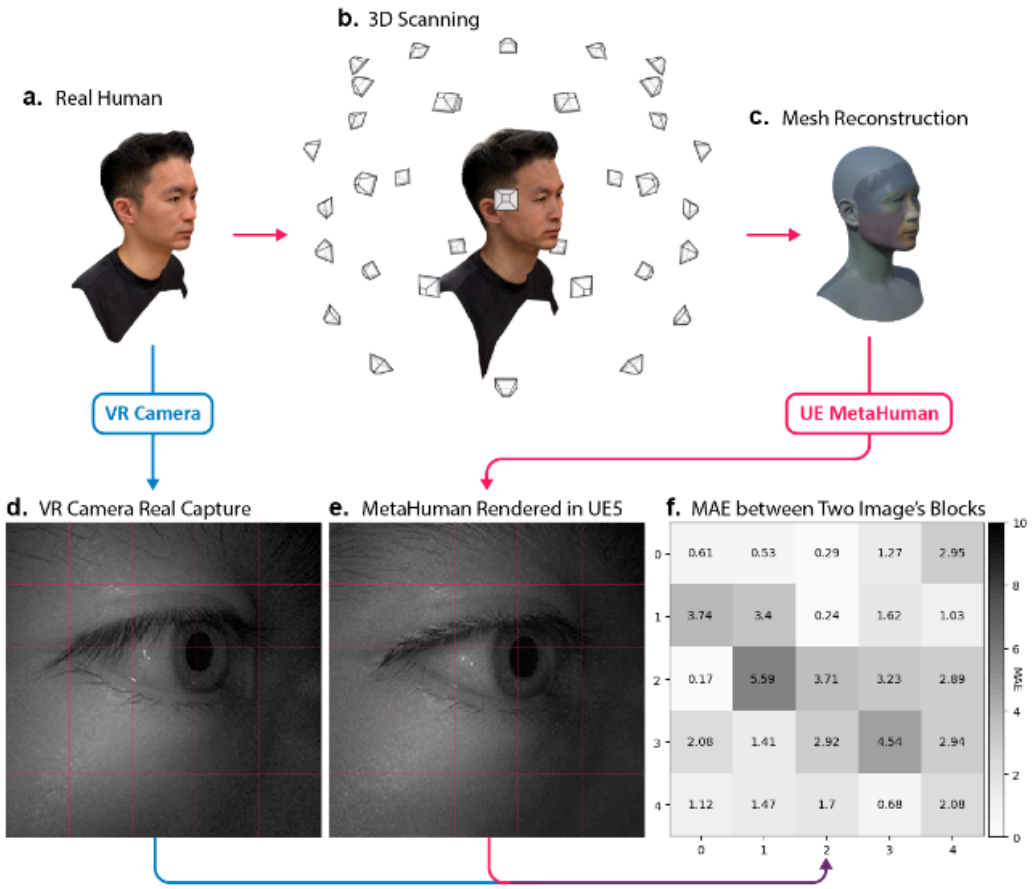


Figure 3: Process of human scan to Metahuman, and synthetic periocular image error quantification and optimisation. **a:** Real human. **b:** 3D reconstruction of the face through RealityCapture. **c:** Converting scanned face mesh to Metahuman. **d:** Image of the periocular taken with real VR headset. **e:** Synthetic periocular image through DPDG environment. **f:**A MAE algorithm iteratively to fine-tune the parameters of the UE cine-camera and UE PointLight to minimise the differences between real and synthetic periocular image.

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

Demographic Distribution After creating the MetaHumans, it's important to describe the demographic distribution among the original pool of 120, which includes characteristics such as gender, age, skin tone, and ethnicity. The diversified representation within the dataset enhances variety and, consequently, bolsters the model's capability to generalise across different scenarios.

In terms of gender distribution, an equitable balance was sought, resulting in an even split of 50% for both male and female MetaHumans. Age representation was approached with inclusivity, accounting for individuals in various life stages including youth (15-24 years), young adults (25-44 years), middle-aged individuals (45-64 years), and the elderly (65 years and above). This approach ensures a broad age range within our data.

The variety of skin tones and ethnic backgrounds was another priority, leading to an inclusive set capturing MetaHumans with varied ethnic backgrounds and skin tones. Specifically, the dataset encompasses individuals of Asian, Caucasian, African, Hispanic backgrounds, as well as other ethnicities (mixed race). Skin tones range from fair to dark, with five categories (Fair, Light Medium, Medium, Medium Dark, Dark) considered in our MetaHuman distribution. It should be noted that due to the constraints of the VR headset capturing only in the IR spectrum, considerations for skin tone did not involve hue, but rather the grayscale intensity under the supplementary infrared light-emitting diodes (IR LEDs) illumination.

Attribute	Category	Count	Percentage
Gender	Male	*60*	*50%*
	Female	*60*	*50%*
Age	Youth (15-24 years)	*26*	*21.67%*
	Young Adult (25-44 years)	*36*	*30%*
	Middle-aged (45-64 years)	*37*	*30.83%*
	Elderly (65 years and above)	*21*	*17.5%*
Skin Tone	Fair	*26*	*21.67%*
	Light Medium	*32*	*26.67%*
	Medium	*30*	*25%*
	Medium Dark	*17*	*14.17%*
	Dark	*15*	*12.5%*
Ethnicity	Asian	*24*	*20%*
	Caucasian	*36*	*30%*
	African	*24*	*20%*
	Hispanic	*24*	*20%*
	Other	*12*	*10%*

Table 1: Distribution characteristics of original 120 MetaHumans.

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

Utilising UE 5.2, we constructed the DPDG environment. Initially, we secured computer-aided design (CAD) models, camera datasets, and parameters of supplementary IR LED from official sources for various VR headsets, including HTC VIVE Pro Eye, HTC Vive Focus 3, Pico 4 Pro, and Varjo XR-3, to precisely align the spatial position and optical parameters of their internal cameras. Figure 4 shows the different distribution of camera positions and corresponding viewpoints in DPDG. Importing CAD models into UE, we placed UE cine-cameras, with parameters adjusted, at the VR internal camera's location. IR LEDs were restored through UE PointLight and were eventually made into Actor blueprints *(Engine, 2023)*. A PostProcessMaterial was applied to the cine-cameras to convert RGB image to depth map image *(Engine, 2023b)*. To streamline the capture process and boost user-friendliness, we composed an automated collection program based on the blueprint using the UE Editor Utility Widget *(Engine, 2023a)*. The program sets the exporting image resolution, selects the VR headset model for collection, sequentially positions the 1150 avatars in the scene, and allows the simulated internal VR camera to individually capture the RGB image and corresponding depth map of each eye. Consequently, 2300 pairs of RGB periocular images and corresponding depth maps per VR headset model were collected.

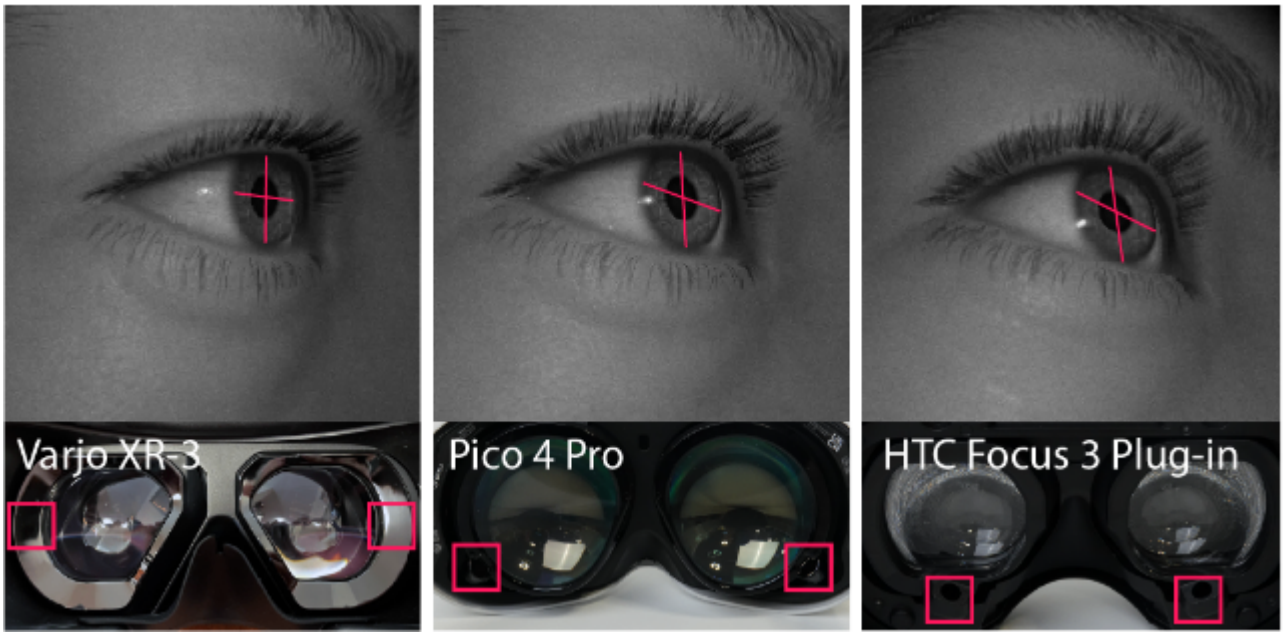


Figure 4: Distribution of monocular camera positions of three VR headsets and corresponding views in DPDG environment. The red cross indicates the perspective reference of the eyeball.

Addressing Robustness Recognising that the unquantified pixel colour disparities between the actual VR capture and the simulation could undermine the robustness of depth map prediction, Despite the highly accurate simulations rendered by MetaHuman, we incorporated a Mean Absolute Error (MAE) metric to quantify differences. This algorithm is used iteratively to fine-tune the parameters of the UE cine-camera and UE PointLight to minimise the differences, thereby optimising the rendering authenticity of the virtual human model.

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

Error Quantification and Optimisation Specifically, we segment each image into $n \times n$ blocks, calculate the average pixel value for each block, and subsequently calculate the MAE between these averages. This block-based approach, rather than calculating the MAE pixel by pixel, mitigates the adverse impact of random noise and minor inconsistencies in facial feature positions on the results, focusing more on the accuracy of the overall colour transition between blocks.

Initially, we divide the actual image I_r and the simulated image I_s into N blocks. For each block i, we compute the average pixel values Avg_{r_i} for the real image and Avg_{s_i} for the simulated image as shown in the equation below:

$$Avg_i = \frac{1}{|B_i|} \sum_{x \in B_i} I_i(x) \quad \text{for} \quad i \in r, s \tag{1}$$

Here, B_i represents the pixel count in block i, while $I_{r_i}(x)$ and $I_{s_i}(x)$ denote the values of pixel x in the real and simulated images in block i respectively.

The MAE for each block, denoted as MAE_i, is computed as the absolute difference between the average pixel values of the reference and the simulated image blocks, specifically Avg_{r_i} and Avg_{s_i}, respectively. Following this, we aggregate the MAEs for all the blocks to derive the total MAE, represented as MAE_{total}. This total MAE is an average of all individual block MAEs and is calculated using the following equation:

$$MAE_{total} = \frac{1}{N} \sum_{i=1}^{N} |Avg_{r_i} - Avg_{s_i}| \tag{2}$$

Here, 'N' signifies the total number of blocks under consideration.

Lastly, using a gradient descent approach, we adjust the parameters of the UE cine-camera's random noise (θ_{noise}) and the UE PointLight's intensity (θ_{light}) to minimise the overall MAE. In each iteration, we update these parameters as a vector $\theta = [\theta_{\text{noise}}, \theta_{\text{light}}]^T$ based on the gradient of the loss function $L(\theta)$, which corresponds to MAE_{total}, as expressed in the equation below:

$$\begin{bmatrix} \theta_{\text{noise}} \\ \theta_{\text{light}} \end{bmatrix} = \begin{bmatrix} \theta_{\text{noise}} \\ \theta_{\text{light}} \end{bmatrix} - \alpha \nabla L \left(\begin{bmatrix} \theta_{\text{noise}} \\ \theta_{\text{light}} \end{bmatrix} \right) \tag{3}$$

Here, α is the learning rate, governing the magnitude of parameter updates at each step. We continue with this update procedure until $L(\theta)$ reaches a minimum value that meets our threshold (0.5%), or the predetermined number of iterations is reached (is set to 100 steps).

Error Analysis To affirm our algorithm's efficacy, we undertook iterative optimisation using four different VR headsets encompassed within our dataset. A representation of the terminal optimisation utilising the Varjo headset is depicted in Figure 3 d-f. Our algorithm consistently

reduced disparities between simulated and actual images, keeping the MAE below an acceptable threshold (average within 1% of the 0-255 pixel colour spectrum), affirming its effectiveness.

Given the depth map range spans from 20mm to 90mm, a 1% error corresponds to an average deviation of 0.7mm in depth estimation under optimal conditions. This degree of error is deemed acceptable for numerous measurement applications within VR. Task-specific errors will be addressed within the Section 4.

Limitations It is important to recognise that the current iteration of MetaHuman is incapable of accurately reconstructing eyebrow and eyelash features, which accounts for why the MAE values of certain blocks are comparatively high. Nonetheless, the overall performance of our optimisation approach offers a robust basis for periocular depth map estimation.

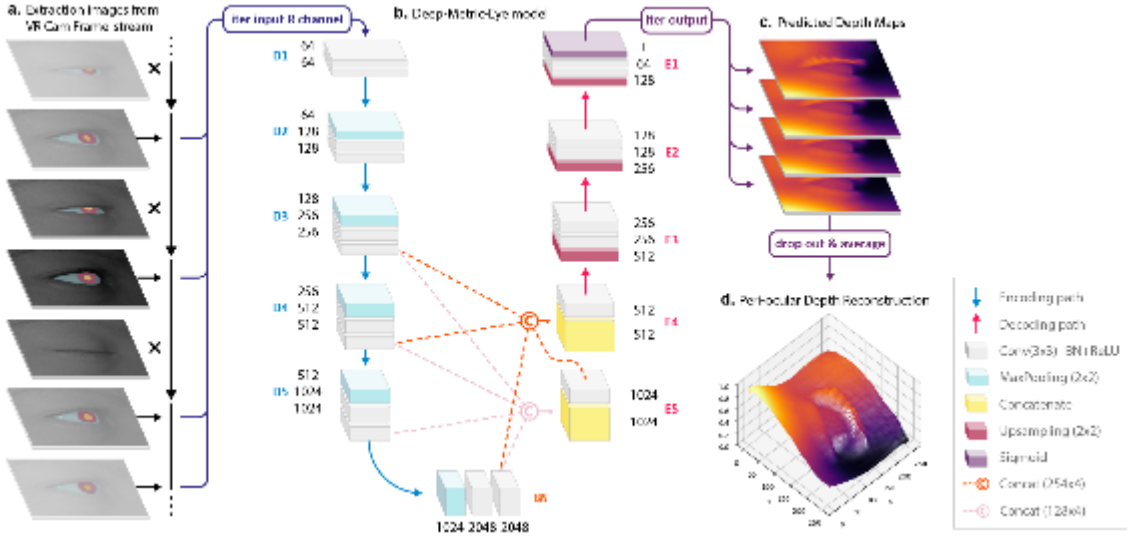


Figure 5: Flowchart of proposed depth estimation framework. **a:** Initial phase involves detection of open-eye state and gaze direction using VR headset's API, from which a sequence of periocular images consistent with open-eye position and gaze direction is extracted from the video stream. **b:** The red channel of extracted images are iteratively input to the depth estimation model, an lightweighted and optimised U-Net 3+ variant with a 5-layer symmetrical encoder-decoder structure. The model omits shallow dense skip connections to diminish the negative impact of intricate details, such as pupils, eyelashes, and eyebrow regions, on the smooth transitions of the depth map, thereby prioritising deep semantics. The numbers indicate the depth dimensions of the tensors. **c:** The output depth maps undergo a two-standard-deviation outlier elimination and pixel averaging to produce **d**, the final periocular depth estimation.

3.2 Model Development

Model Overview A three-stage periocular depth estimation model is proposed in this study. Figure 5 illustrates the framework. In the initial stage, we utilise the VR headset's eye edge detection API to extract a series of images from the video frame flow, ensuring they align with specific eye-opening conditions. Subsequently, the selected images are iteratively input into a deep learning network, which is composed of a 5-layer symmetrical encoder-decoder, for the generation of corresponding predicted depth maps. Finally, these depth maps undergo outlier exclusion via a two-standard-deviation threshold and averaging to produce the ultimate periocular depth

prediction. Further application of this model, in conjunction with the eye feature recognition API of the VR headset, allows for the calculation of metric dimensions of arbitrary features.

3.2.1 Data Pre-processing

Multi-image Collection Strategy Given the nature of wearing a VR headset, the relative position of the eye and the camera remains constant. In periocular depth estimation tasks, the process is executed as a long interval operation where accuracy of the single operation is given priority over high frame-rate processing. Therefore, we design a multi-image collection strategy for the data preparation phase. This strategy utilises built-in eye feature segmentation APIs of various VR systems to judge and collect multiple images consistent in eyelid opening degree and gaze direction. Once multiple estimated depth maps are obtained, outlier detection and processing are performed using the Median Absolute Deviation (MAD) method. MAD, a robust outlier detection method based on median absolute deviation, is less prone to extreme value impact than the standard Z-score *(Curtis et al., 2016)*. Utilising this method effectively eliminates outliers in depth maps, and averages are taken for each pixel to improve prediction robustness.

VR API Integration All current VR headsets with eye-oriented cameras feature eye segmentation APIs for pupil and gaze detection. Given user privacy concerns, these APIs are proprietary. We obtained testing permissions through NDAs with HTC and PICO. For Varjo XR-3 PC VR, direct access to eye images captured by the camera is available. To ensure the complete reproducibility of the entire process, the open-source human eye segmentation framework, EllSeg, proposed by Kothari et al., was also employed as a direct-use implementation for Varjo XR-3 *(Kothari et al., 2021)*.

Image Capture Condition The extent of eyelid opening is crucial for tasks like facial expression recognition, fatigue detection, and depth estimation. Our method for determining eye openness is a cost-efficient computational strategy, which doesn't necessitate a specific detection orientation and has been robustly tested within systems by Meta and HTC. The approach begins by segmenting and curve-fitting the upper and lower eyelid contours. Subsequently, it identifies the midpoint of both curves, evaluating the amplitude intervals and extremities of their movement to deduce the level of eye openness. In our study, images are captured when the gaze direction is straight ahead and the eyelids are naturally at their widest extent.

Model Consideration Transitioning from data preparation, our key challenge was designing a deep learning model suitable for the VR context, especially for standalone headsets with their computational and memory constraints. We considered architectures like DenseNet, ResNet, MonoDepth, Transformer, and Deeper Depth Prediction, all of which are renowned for depth estimation tasks*(Godard et al., 2017; He et al., 2016; Iandola et al., 2014; Laina et al., 2016; Vaswani et al., 2017)*. While these models have demonstrated significant performance in their respective domains, their inherent complexity and high parameter count make them less suitable for the constrained resources of VR applications. For instance, DenseNet's dense connections between all layers and ResNet's multiple stacked residual blocks, although beneficial for their specific use

cases, escalate the model complexity and parameter count, posing challenges in the VR context. Increased parameters entail augmented memory and computational demands, a challenge for standalone VR headsets bereft of potent GPUs. Moreover, the inherent architectural complexity augments energy consumption, potentially hastening battery depletion and inducing latency disruptions in real-time interactions.

Re-imagined U-Net Architecture for VR In the evolution from PC VR to standalone VR systems, there's a pronounced need for a lightweight yet robust deep learning model. The original U-Net architecture, commendable in its own right, faced challenges. Specifically, the classic 4-layer encoder-decoder network struggled to adequately abstract conflicting colour features, such as hair and pupils, as can be observed in the details of Figure 6. To address this, we reconfigured the U-Net blueprint. We judiciously omitted shallow skip connections, which, although beneficial for certain applications, introduced potential noise due to their emphasis on low-level features. Instead, we favoured a system that separately handles abstraction and concrete expressiveness. Further, in the deeper layers of the network, weights were assigned based on the depth of the network to each encoding layer participating in skip connections, controlling the transmission of feature details and abstraction. By augmenting the classic 4-layer structure with an additional layer, transforming it into a 5-layer symmetric structure, our model bolstered its capacity to accurately capture the nuanced depth variations inherent to the eye region. While PC-based depth estimation solutions often veer towards increased complexity, our re-imagined U-Net strikes a balance between efficiency and performance, offering a promising approach for similar challenges, especially in the context of resource-constrained VR devices.

3.2.2 Model Training

Training Dataset Preparation Model training was performed separately for each VR headset model. For each model, 2300 pairs of periocular RGB images and corresponding ground truth depth maps, both of 256x256 resolution, were randomly apportioned into training, validation, and testing subsets, adhering to a 70%-15%-15% distribution. As the camera within the VR is equipped with a filter film that matches the wavelength of the supplementary IR LEDs to avoid screen light interference, it results in its effective light-sensitive channel being the red channel only. Therefore, during the data pre-processing stage, in order to be consistent with the VR camera and to further reduce the number of model parameters, the RGB images used for training were channel split, with only the red single channel data being extracted for training and the green and blue channels being discarded.

Model Architecture Detail As shown in Figure 5 b, our model comprises a five-layer down-sampling convolutional encoder, a bottleneck layer, and a five-layer up-sampling convolutional decoder. Each layer of the five-layer encoder and the bottleneck layer employs a convolution operation with a 3x3 kernel, stride of 2, and padding of 1, followed by a ReLU activation function and batch normalisation. This sequence is repeated once within each layer. Pooling operations with a pool size and stride of 2 succeed each encoder layer, further reducing the spatial dimensions.

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

In the decoder part, we introduce skip connections to enable feature-sharing from the third, fourth, fifth layer of encoder and bottleneck layer with the corresponding fourth and fifth layer in the decoder. Additionally, to regulate the propagation of features during decoding, we assign varying weights to the encoder layers based on their depth. This strategic weighting ensures that our model captures and propagates the necessary level of detail during the encoding-decoding process. The specific weights applied are represented by the following function:

```
def skip_concatenate(E3, E4, E5, BN):
    """
    Weights Set for Model Skip Concatenations
    """
    # Level-5 decoder feature map
    D5 = concat(
        0.1 * E3,
        0.8 * E4,
        1.0 * E5,
        1.0 * BN
    )

    # Level-4 decoder feature map
    D4 = concat(
        0.2 * E3,
        0.5 * E4,
        0.8 * BN,
        1.0 * D5
    )

    return D4, D5
```

The third to fifth decoder layers initially undergo a bilinear up-sampling operation with a scale of 2, succeeded by two sets of 3x3 convolutions (stride 2, padding 1), ReLU activation, and batch normalisation. Finally, we apply a 1x1 convolution, followed by a Sigmoid activation function. All convolutional layers adopt He initialisation for weights and zero initialisation for biases.

Loss Function The chosen loss function for our depth estimation task is the Reverse Huber Loss (BerHu). The BerHu loss function, given by Equation 4, exhibits different behaviours contingent on the magnitude of the error, (x), which corresponds to the disparity between the predicted and actual depth values. The parameter (c) denotes a threshold, which has been set at 24% of the maximum absolute error within a mini batch after our iterative testing.

$$L_{\text{berHu}}(x) = \begin{cases} |x| & \text{if } |x| \leq c, \\ \frac{x^2+c^2}{2c} & \text{otherwise,} \end{cases} \tag{4}$$

For errors that do not exceed c, BerHu aligns with the L1 loss function, thereby functioning as an absolute error. This linear component renders the loss function resilient to outliers, mitigating their impact on the model's learning trajectory and thereby enhancing model robustness.

In the scenario where errors exceed c, BerHu mirrors the L2 loss function, acting as a squared error. This quadratic aspect imposes more substantial penalties on larger errors. The consequential drive to minimise these errors results in the model paying enhanced attention to such instances.

The distinct dual character of the BerHu loss function enables effective management of both minor and major prediction errors, demonstrating its suitability for depth estimation tasks.

Regularisation We adopt several strategies to counteract over-fitting. Dropout layers, with a dropout rate of 0.5, are inserted into the fifth layer of the encoder and the bottleneck layer. Additionally, we introduce an early stopping mechanism that halts training if there is no improvement in the validation loss after 20 consecutive epochs. Training is set for a maximum of 150 epochs, starting with an initial learning rate of 1e-4.

Model	AbsRel ↓	SqRel ↓	RMSE ↓	RMSElog ↓	δ<1.25 ↑	δ<1.25² ↑	δ<1.25³ ↑	Parameters ↓
Ours-VIVE Pro Eye	0.039	0.012	0.013	1.009	0.979	0.985	0.986	28.8 M
Ours-VIVE Focus3	0.041	0.014	0.021	0.921	0.971	0.981	0.983	28.8 M
Ours-Pico 4 Pro	0.037	0.012	0.018	1.109	0.969	0.986	0.982	28.8 M
Ours-Varjo XR-3	0.035	0.011	0.014	1.122	0.976	0.984	0.987	28.8 M
Ours Avg.	**0.038**	**0.012**	**0.017**	**1.040**	**0.974**	**0.984**	**0.985**	28.8 M
U-Net 4|4-layer	0.648	0.161	0.182	1.830	0.244	0.558	0.818	31.1 M
U-Net 5|5-layer	0.064	0.021	0.022	1.155	0.963	0.982	0.978	53.6 M
U-Net3+ 4|4-layer	0.121	0.029	0.043	1.119	0.957	0.980	0.984	**26.9 M**
U-Net3+ 5|5-layer	0.092	0.018	0.038	1.023	0.965	0.983	0.979	44.6 M

Table 2: Comparative analysis of our model and the original structured U-Net models in terms of depth error, depth accuracy, and parameters. The depth error and parameters are preferred to be lower (represented by ↓), and depth accuracy is preferred to be higher (represented by ↑). Results are given for different VR headset datasets and the average performance is also provided. Our model consistently outperforms the U-Net models, showing superior performance across different headset datasets and maintaining a comparable model size to the 4-layer U-Net3+ model.

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

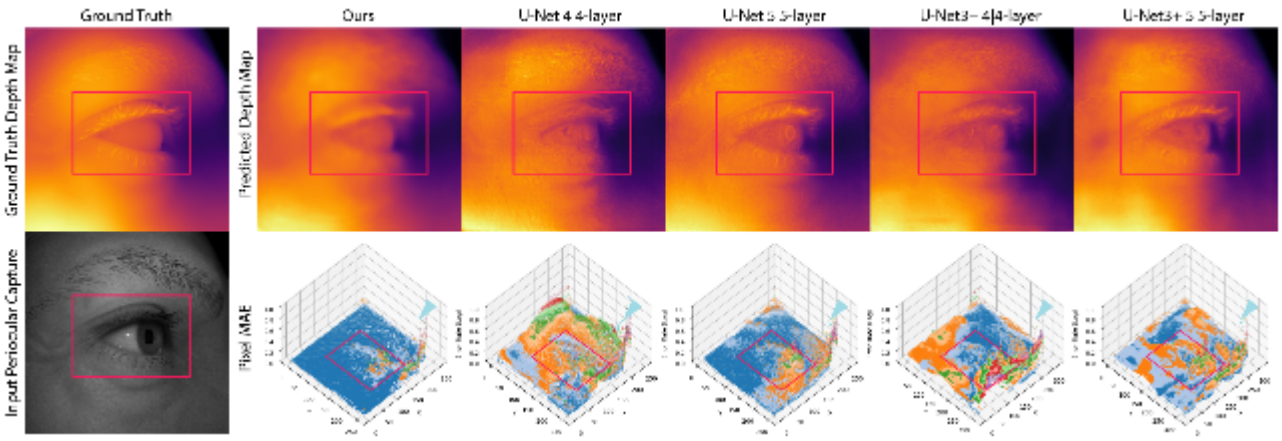


Figure 6: A three-dimensional visualisation of the MAE for the depth estimation from our model and four original structured U-Net models. The MAE values have been log-transformed to emphasise the differences. The x and y-axes represent the length and width of the image, while the z-axis represents the log-transformed MAE values. The visualisation demonstrates our model's superior depth estimation performance, particularly in intricate areas such as around the pupils and cheekbones. It highlights the benefits of our model's focus on deep semantics, and the shortcomings of models giving equal attention to both surface details and deeper semantics.

3.2.3 Model Evaluation

Evaluation and Comparative Analysis We synthesised perspective-appropriate training datasets for various VR headsets in the DPDG environment and subjected our model to training and evaluation. To present an integrated perspective on the model's performance, we averaged the model performance metrics across different headsets and compared them with four original structured U-Net models as the benchmark. These encompass the U-Net and U-Net 3+ models with both 4-layer and 5-layer symmetrical structures.

Multiple metrics were employed to measure depth estimation error (lower is better) and accuracy (higher is better), including Absolute Relative Difference (AbsRel), Squared Relative Difference (SqRel), Root Mean Square Error (RMSE), Logarithmic Root Mean Square Error (RMSElog), and the accuracy under thresholds of $\mathbf{1.25}$, $\mathbf{1.25^2}$, and $\mathbf{1.25^3}$.

The results, summarised in Table Table 2, clearly demonstrate that, compared to the four original U-Net models, our model consistently outperforms in terms of both depth estimation error and accuracy across all tested VR headsets, indicating substantial generalisability. Moreover, it is worth noting that our model achieves a deeper layer structure with a parameter count comparable to that of the 4-layer symmetrical U-Net 3+ model, proving to be more efficient than the other three models.

Visualisation of Errors For a detailed performance insight, we visualised depth estimation results from a representative image in the validation set Figure 6. This involved comparing our model's estimations with the ground truth and calculating the MAE for each normalised depth value. The log-transformed MAE values, shown in a 3D plot, indicate our model's superior accuracy, especially around the pupils and cheekbones, suggesting its utility for tasks like measuring pupil diameter. In

comparison, other models struggle with regions like the pupil due to their equal focus on surface and deeper semantics.

Analysis Summary On the whole, our model manifests a robust performance. In comparison with the original U-Net models, our model demonstrates superior depth estimation accuracy and lower error across multiple VR headsets. Moreover, the visualisation further emphasises the advantages of our model, especially in estimating the depth of complex regions of detail and transition such as eyelashes, eyebrows, and image boundaries.

4 Experiment

To demonstrate our VR depth estimation model's practical applicability, we conducted two experiments. In the **Global Precision Evaluation**, we employed the ORBBEC Femto Time of Flight (ToF) camera, chosen for its renowned near-limit accuracy in commercial devices, boasting a depth map accuracy error of only 0.2%. By comparing the depth data from participants' periocular regions captured by this ToF camera with our model's estimations, this experiment aimed to evaluate our model's global region error, assessing its potential as a feasible alternative to high-precision equipment. The **Pupil Diameter Measurement** experiment evaluated the model's capability to measure fine details. Under constant lighting, we measured participants' pupil sizes and contrasted them with sizes from our model's depth maps. Additionally, prediction time for each depth map was recorded. Figure 7 depicts the experimental setup.

36 participants (18-60 years, balanced by gender) with a variety of skin tones participated in our study. They were health-screened, with no vision impairments. To ensure uniformity, glasses and contact lenses were removed, and of these participants, 12 wore makeup. All participants gave informed consent after understanding the study's procedures and potential discomforts. They were informed of their right to withdraw anytime. The study was approved by the Institutional Review Board of Royal College of Art, adhering to ethical standards.

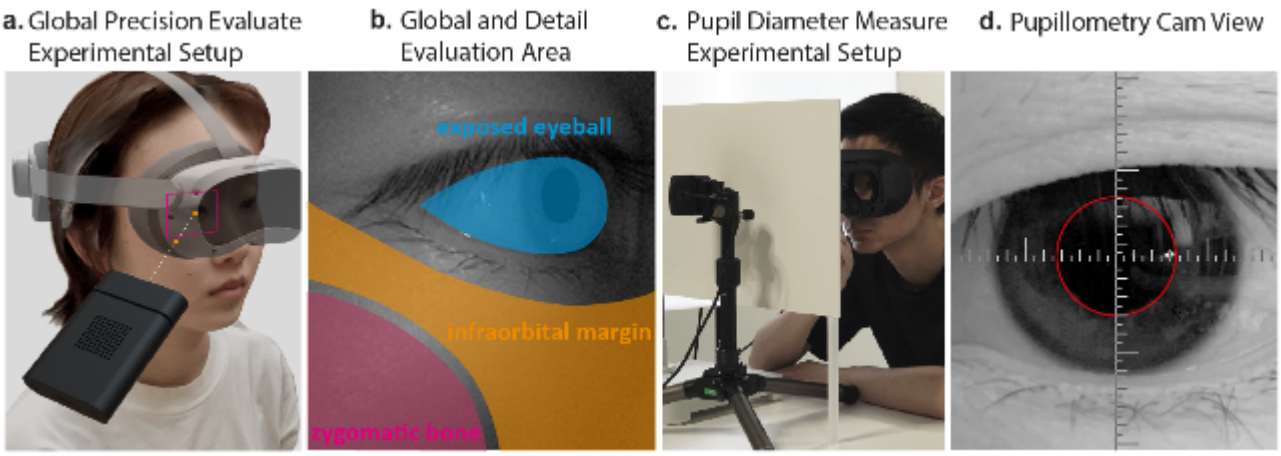


Figure 7: Experimental Setup **a:** ORBBEC Femto ToF camera and PICO 4 Pro VR headset alignment for periocular region imaging. **b:** Illustration of the regions used for global and regional accuracy evaluation. **c:** Setup for the pupil measurement experiment. **d:**Pupillometry Camera view for pupil diameter measurement.

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

4.1 Global Precision Evaluation

In the first experiment, each participant underwent two periocular data acquisitions for both the left and right eyes. Initially, participants were equipped with a PICO 4 Pro VR headset. We deployed our proposed model within the VR system using the *Pytorch Android Mobile* deployment process to estimate depth maps at a 256x256 pixel resolution. To establish a consistent baseline and counteract camera distortion, we transformed the depth map to absolute spatial coordinates using the equation:

$$(X, Y, Z) = ((x - c_x) * s * D_p / f_x, (y - c_y) * s * D_p / f_y, s * D_p) \quad (5)$$

where D_p represents the pixel value in the depth map, while s is a conversion scaling factor used to convert pixel values into the real-world depth, Z. The coordinates x and y represent the pixel location in the image, and (X, Y, Z) are the corresponding coordinates in three-dimensional space. c_x and c_y are the positions of the optical centre of the camera lens in the image coordinate system, typically located at the centre of the image. The terms f_x and f_y denote the focal lengths of the camera.

For the second acquisition, the Femto ToF camera was aligned to mirror the viewpoint of the PICO headset's internal camera and was positioned 20cm away from the participant's eye (its minimum effective focal distance), capturing periocular depth data at a 256x256 pixel resolution. The absolute coordinates were derived from the official camera API.

Upon concluding the measurements, we amassed 72 valid samples. The pixel-based Mean Absolute Error (MAE_p) between the two depth maps was calculated to be 1.68mm, with a standard deviation (sd) of 1.22mm across all pixels. The predominant errors, mainly observed in the eyebrow and eyelash regions, arise from the Femto camera's limitations in effectively capturing hair. Additionally, our model's strategic abstraction of these regions to achieve surface smoothness might also be a contributing factor to the discrepancies.

However, when focusing on regions of paramount importance for medical standards, such as the exposed eyeball, infraorbital margin, and zygomatic bone, our model showcased commendable precision. The MAE_p values for these regions were 0.63mm, 0.74mm, and 0.57mm, respectively, with sd values of 0.27mm, 0.35mm, and 0.24mm. These findings underline the potential of our approach in practical, medically-relevant scenarios. Notably, despite the makeup variations across participants, the results remained consistent, affirming the robustness of our model, which was trained on datasets inclusive of diverse makeup levels. On average, the model required 8.11 seconds to predict a depth map.

Full | DOI:1234.5678 | Vol.1 No.1 (2025)

4.2 Pupil Diameter Measurement

For the secondary experiment, participants were positioned in a controlled environment with consistent illumination, oriented towards a uniformly frosted cardboard sheet. Centrally, an aperture was incorporated into the cardboard to facilitate a calibrated *Baumer VCXU.2-123C* pupillometry camera, purposed for frontally gauging participants' pupil diameters. Concurrently, participants were equipped solely with the HTC Focus 3 eye tracker. The main body of the VR headset, tethered via a USB type-c cable, was strategically placed laterally. Depth map prediction was executed via the *Pytorch Android Mobile* deployment within the VR system. The pupillometry camera and HTC eye tracker accessory operated in tandem, producing reference diameter measurements and predictive depth maps. Following this, the absolute pupil diameter was ascertained employing the VR system's intrinsic pupil segmentation API in conjunction with equation Equation 5. The resulting discrepancy was quantified against the reference value by computing the absolute differential between the ground truth pupil size and the size deduced from the depth map.

Notably, while corneal refraction possesses the potential to perturb pupil diameter measurements, preliminary simulations, constructed via Zemax modelling of an idealised human ocular structure, suggest that refractive aberrations owing to corneal interference are inconsequential under our experimental conditions, especially when assessing pupil diameter at oblique angles.

Subsequent to experimental completion, a corpus of 72 valid measurement samples was curated, delineating pupil diameters spanning from 3.61mm to 4.48mm, averaging at 5.17mm. The divergence between authentic and estimated pupil diameters yielded an average (μ) of 0.33mm, accompanied by a standard deviation (sd) of 0.14mm. The ensuing percentage discrepancy totaled 6.38%. The mean computational time required for depth map prediction was registered at 7.31 seconds.

4.3 Analysis of Results

The experimental outcomes validate the efficacy and precision of our periocular depth estimation framework in both feature measurement and micro-scale feature assessment. In the Global Precision Evaluation experiment, the model showcased high accuracy in regions critical for health standards calculation, despite diminished precision in areas covered by hair. However, in the Pupil Diameter Measurement experiment, a slight increase in error was observed. We hypothesise this discrepancy results from the limited pixel area occupied by the pupil in low-resolution images, potentially leading to imprecise pupil edge segmentation due to pixel aliasing. Addressing this limitation by enhancing acquisition resolution or employing curve fitting could refine the accuracy of micro-scale feature estimations. These findings underscore the framework's viability for advanced eye state monitoring in VR settings, aligning with light stimulation standards and medical guidelines.

5 Conclusion

In summary, our study presents a composite framework for periocular measurable depth estimation capable of efficiently and accurately predicting spatial metric dimensions for eye region features using an eye-oriented monocular VR camera. To mitigate the challenges of facial data collection for training the model, we introduce our DPDG environment, which can generate synthetic periocular datasets for various VR headsets using MetaHuman. Through having conducted two practical experiments, the Global Precision Evaluation and the Pupil Diameter Measurement, our model has proven its robust capabilities in spatial depth estimation and detailed small-scale feature assessment. Our aim is to bridge the gap between eye and periocular state changes during VR immersion and light stimulation standards and medical guidelines, thereby facilitating deep and comprehensive monitoring, with the ultimate goal of effectively quantifying stimuli to mitigate the harm inflicted on human eyes by current VR headsets usage.

Limitation and Future Work Despite the successful deployment of our framework in VR, the volume of parameters remains substantial. We hypothesise that this could be due to the profusion of superfluous details in the eyelash and eyebrow areas, which may have significantly consumed network fitting efficiency and drastically increased the parameter count. To address this, a potentially effective solution might involve deleting eyelashes and eyebrows when creating ground truth depth maps within the DPDG, while preserving the details in the RGB images for model training. Another limitation is the exclusive focus on participants not wearing glasses. The presence of glasses could pose heightened challenges in depth estimation due to reflections, lens distortions, and obstructions, warranting exploration in future studies. Although the current straight-forward gaze acquisition suffices for most scenarios, we envision potential dynamic acquisition in the future by integrating 3D ocular surface reconstruction*(Dierkes et al., 2018; Yiu et al., 2019)*. Lastly, While pixel-based edge segmentation offers speed advantages, it can result in jagged edges when measuring small-scale features such as the pupil, leading to reduced accuracy. Transitioning to curve fitting might provide a more refined approach, enhancing precision.

6 Open Science

To promote further research into eye health within VR, we have made our DPDG dataset synthesis environment and the depth estimation model code available to the public through the provided GitHub link. For the experimental metadata containing facial information, please contact the authors and sign an NDA before accessing, to ensure the facial privacy rights of the participants.

References

Chaudhary, A. K., Kothari, R., Acharya, M., Dangi, S., Nair, N., Bailey, R., Kanan, C., Diaz, G., & Pelz, J. B. (2019). Ritnet: Real-time semantic segmentation of the eye for gaze tracking. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 3698–3702.

Chen, X., Chen, X., & Zha, Z.-J. (2019). Structure-aware residual pyramid network for monocular depth estimation. *arXiv Preprint arXiv:1907.06023*.

Chen, Y., Wang, X., & Xu, H. (2021). Human factors/ergonomics evaluation for virtual reality headsets: a review. *CCF Transactions on Pervasive Computing and Interaction*, *3*(2), 99–111.

Cheng, L., Chen, S., Liu, X., Xu, H., Wu, Y., Li, M., & Chen, Y. (2018). Registration of laser scanning point clouds: A review. *Sensors*, *18*(5), 1641.

Curtis, A. E., Smith, T. A., Ziganshin, B. A., & Elefteriades, J. A. (2016). The mystery of the Z-score. *Aorta*, *4*(04), 124–130.

Dierkes, K., Kassner, M., & Bulling, A. (2018). A novel approach to single camera, glint-free 3D eye model fitting including corneal refraction. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 1–9.

Dincelli, E., & Yayla, A. (2022). Immersive virtual reality in the age of the Metaverse: A hybrid-narrative review based on the technology affordance perspective. *The Journal of Strategic Information Systems*, *31*(2), 101717.

Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, *27*.

Engine, U. (2023a). Editor utility widgets, https://docs.unrealengine.com/5.2/en-US/editor-utility-widgets-in-unreal-engine/. In *Unreal Engine 5.2 Documentation - Editor Utility Widgets*. https://docs.unrealengine.com/5.2/en-US/editor-utility-widgets-in-unreal-engine/

Engine, U. (2023b). Point lights, https://docs.unrealengine.com/5.2/en-US/point-lights-in-unreal-engine/. In *Unreal Engine 5.2 Documentation - Point Lights*. https://docs.unrealengine.com/5.2/en-US/point-lights-in-unreal-engine/

Engine, U. (2023c). Post process materials, https://docs.unrealengine.com/5.2/en-US/post-process-materials-in-unreal-engine/. In *in Unreal Engine | Unreal Engine 5.2 Documentation*. https://docs.unrealengine.com/5.2/en-US/post-process-materials-in-unreal-engine/

Fang, Z., Cai, L., & Wang, G. (2021). MetaHuman Creator The starting point of the metaverse. *2021 International Symposium on Computer Technology and Information Science (ISCTIS)*, 154–157.

Feder, R. S., Olsen, T. W., Prum, B. E., Summers, C. G., Olson, R. J., Williams, R. D., & Musch, D. C. (2016). Comprehensive adult medical eye evaluation Preferred Practice Pattern® guidelines. *Ophthalmology*, *123*(1), P209–P236.

Fuhl, W., Geisler, D., Rosenstiel, W., & Kasneci, E. (2019). The applicability of Cycle GANs for pupil and eyelid segmentation, data generation and image refinement. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Fuhl, W., Kasneci, G., & Kasneci, E. (2021). Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 367–375.

Fuhl, W., Weber, D., & Eivazi, S. (2022). Pistol: Pupil invisible supportive tool to extract pupil, iris, eye opening, eye movements, pupil and iris gaze vector, and 2d as well as 3d gaze. *arXiv Preprint arXiv:2201.06799*.

Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 270–279.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hirzle, T., Fischbach, F., Karlbauer, J., Jansen, P., Gugenheimer, J., Rukzio, E., & Bulling, A. (2022). Understanding, addressing, and analysing digital eye strain in virtual reality head-mounted displays. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *29*(4), 1–80.

Holmqvist, K., Örbom, S. L., Hooge, I. T., Niehorster, D. C., Alexander, R. G., Andersson, R., Benjamins, J. S., Blignaut, P., Brouwer, A.-M., Chuang, L. L., & others. (2023). Eye tracking: empirical foundations for a minimal reporting guideline. *Behavior Research Methods*, *55*(1), 364–416.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., & Wu, J. (2020). Unet 3+: A full-scale connected unet for medical image segmentation. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1055–1059.

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., & Keutzer, K. (2014). Densenet: Implementing efficient convnet descriptor pyramids. *arXiv Preprint arXiv:1404.1869*.

Jin, Y., Yu, L., Chen, Z., & Fei, S. (2021). A Mono SLAM Method Based on Depth Estimation by DenseNet-CNN. *IEEE Sensors Journal*, *22*(3), 2447–2455.

Kang, J., Han, X., Song, J., Niu, Z., & Li, X. (2020). The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data. *Computers in Biology and Medicine*, *120*, 103722.

Kansal, P., & Devanathan, S. (2019). Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 3688–3693.

Keshavarz, B., & Golding, J. F. (2022). Motion sickness: current concepts and management. *Current Opinion in Neurology*, *35*(1), 107–112.

Kothari, R. S., Chaudhary, A. K., Bailey, R. J., Pelz, J. B., & Diaz, G. J. (2021). Ellseg: An ellipse segmentation framework for robust gaze tracking. *IEEE Transactions on Visualization and Computer Graphics*, *27*(5), 2757–2767.

Laga, H., Jospin, L. V., Boussaid, F., & Bennamoun, M. (2020). A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(4), 1738–1764.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. *2016 Fourth International Conference on 3D Vision (3DV)*, 239–248.

Li, F., Lee, C.-H., Feng, S., Trappey, A., & Gilani, F. (2021). Prospective on eye-tracking-based studies in immersive virtual reality. *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 861–866.

Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F. X., Taylor, R. H., & Unberath, M. (2021). Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6197–6206.

Lim, J. Z., Mountstephens, J., & Teo, J. (2020). Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors*, *20*(8), 2384.

Lisk, S., Vaswani, A., Linetzky, M., Bar-Haim, Y., & Lau, J. Y. (2020). Systematic review and meta-analysis: Eye-tracking of attention to threat in child and adolescent anxiety. *Journal of the American Academy of Child & Adolescent Psychiatry*, *59*(1), 88–99.

Mehrfard, A., Fotouhi, J., Taylor, G., Forster, T., Navab, N., & Fuerst, B. (2019). A comparative analysis of virtual reality head-mounted display systems. *arXiv Preprint arXiv:1912.02913*.

Min-Allah, N., Jan, F., & Alrashed, S. (2021). Pupil detection schemes in human eye: a review. *Multimedia Systems*, *27*(4), 753–777.

Ming, Y., Meng, X., Fan, C., & Yu, H. (2021). Deep learning for monocular depth estimation: A review. *Neurocomputing*, *438*, 14–33.

Modi, N., & Singh, J. (2021). A review of various state of art eye gaze estimation techniques. *Advances in Computational Intelligence and Communication Technology: Proceedings of CICT 2019*, 501–510.

Mystakidis, S. (2022). Metaverse. *Encyclopedia*, *2*(1), 486–497.

Ning, H., Wang, H., Lin, Y., Wang, W., Dhelim, S., Farha, F., Ding, J., & Daneshmand, M. (2023). A Survey on the Metaverse: The State-of-the-Art, Technologies, Applications, and Challenges. *IEEE Internet of Things Journal*.

on Illumination, I. C. (2018). CIE system for metrology of optical radiation for iPRGC-influenced responses to light. In *Report CIE S 026/E: 2018* (p. 32). CIE Vienna, Austria.

on Illumination, I. C. (2022). Light and Lighting — Integrating Lighting — Non-visual Effects. In *ISO/CIE TR 21783:2022(E)*. CIE Vienna, Austria.

Reality, C. (2023). Explore the possibilities, https://www.capturingreality.com/realitycapture. In *RealityCapture - Photogrammetry Software*. https://www.capturingreality.com/realitycapture

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241.

RP-3-13, I. E. S. A. (2014). *American National Standard Practice on Lighting for Educational Facilities*. Illuminating Engineering Society of North America New York, NY, USA.

Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *Ieee Access*, *9*, 82031–82057.

Souchet, A. D., Lourdeaux, D., Pagani, A., & Rebenitsch, L. (2023). A narrative review of immersive virtual reality's ergonomics and risks at the workplace: cybersickness, visual fatigue, muscular fatigue, acute stress, and mental overload. *Virtual Reality*, *27*(1), 19–50.

Souchet, A. D., Philippe, S., Lourdeaux, D., & Leroy, L. (2022). Measuring visual fatigue and cognitive load via eye tracking while learning with virtual reality head-mounted displays: A review. *International Journal of Human–Computer Interaction*, *38*(9), 801–824.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, K., & Ji, Q. (2017). Real time eye gaze tracking with 3d deformable eye-face model. *Proceedings of the IEEE International Conference on Computer Vision*, 1003–1011.

Wang, Y., Zhai, G., Chen, S., Min, X., Gao, Z., & Song, X. (2019). Assessment of eye fatigue caused by head-mounted displays using eye-tracking. *Biomedical Engineering Online*, *18*, 1–19.

Wolffsohn, J. S., Lingham, G., Downie, L. E., Huntjens, B., Inomata, T., Jivraj, S., Kobia-Acquah, E., Muntz, A., Mohamed-Noriega, K., Plainis, S., & others. (2023). TFOS Lifestyle: impact of the digital environment on the ocular surface. *The Ocular Surface*, *28*, 213–252.

Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., & Ricci, E. (2018). Structured attention guided convolutional neural fields for monocular depth estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3917–3925.

Yiu, Y.-H., Aboulatta, M., Raiser, T., Ophey, L., Flanagin, V. L., Zu Eulenburg, P., & Ahmadi, S.-A. (2019). DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. *Journal of Neuroscience Methods*, *324*, 108307.

Zanlorensi, L. A., Laroca, R., Lucio, D. R., Santos, L. R., Britto Jr, A. S., & Menotti, D. (2022). A new periocular dataset collected by mobile devices in unconstrained scenarios. *Scientific Reports*, *12*(1), 17989.

Zemblys, R., Niehorster, D. C., & Holmqvist, K. (2019). gazeNet: End-to-end eye-movement event detection with deep neural networks. *Behavior Research Methods*, *51*, 840–864.

Zhang, S. (2018). High-speed 3D shape measurement with structured light methods: A review. *Optics and Lasers in Engineering*, *106*, 119–131.

Zhang, Z., Wen, F., Sun, Z., Guo, X., He, T., & Lee, C. (2022). Artificial intelligence-enabled sensing technologies in the 5G/internet of things era: from virtual reality/augmented reality to the digital twin. *Advanced Intelligent Systems*, *4*(7), 2100228.

Zhao, C., Sun, Q., Zhang, C., Tang, Y., & Qian, F. (2020). Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, *63*(9), 1612–1627.

Zhou, L., & Kaess, M. (2020). Windowed bundle adjustment framework for unsupervised learning of monocular depth estimation with u-net extension and clip loss. *IEEE Robotics and Automation Letters*, *5*(2), 3283–3290.