A. Data Quality

Table 8. **Generators.** We experiment with different image generators [50, 53] and control mechanisms [45, 73]. Our method is not specific to any combination and improves with better generators.

Model	Condition	FID↓	IS↑
SD 1.5	ControlNet	7.86	13.59
SDXL	ControlNet	4.18	**15.09**
SDXL	T2I	**3.22**	14.37

We evaluate the impact of different diffusion models and conditioning mechanisms by generating 20,000 images per configuration. Using FID [23] and Inception Score [54], we measure scene diversity and realism, evaluating full images rather than face crops focusing on scene composition and multi-person layout quality, crucial for robust real-world performance.

Tab. 8 demonstrates that advanced diffusion models [50] significantly outperform earlier versions [53] in generating high-quality multi-person scenes. While ControlNet [73] and T2I-Adapter [45] achieve comparable metrics, ControlNet frequently produces anatomical deformations, leading us to choose T2I-Adapter for more reliable scene generation.

B. Re-Identification on Synthetic Data

Figure 6 provides visual examples of generated samples that were matched with public figures from Celeb-A [41] dataset. Even though the data generation include the prompt anonymization, some parts of the prompts, such as the movie name still might encode identity of famous people. Identifying and removing such samples improve the privacy aspect. Fig. 7 shows an example of an image generated by diffusion model that was finetuned on RaFD [32] dataset, and 2 nearest neighbors from the dataset it was trained on. Even though the model learns features from the original images, in most cases it does combine features from multiple people and fails to recover finer details that encode the identity.

C. Model Details

The architecture is inspired by object detection methods like CenterNet [14] or YOLO-NAS-Pose [2], enabling efficient end-to-end learning. The model is based on the YOLO-NAS [2] architecture for object detection. YOLO-NAS use a neural architecture search engine to enhance the YOLO family of models by optimzing the sizes and structures of stages, block types, the number of blocks, and the number of channels in each stage. We employ the YOLO-NAS-L backbone, though the model is agnostic to the choice of encoder. The "neck" is used to fuse the features generated by the backbone. The visual features from the encoder maps neck are fused by the Spatial Pyramid Pooling [21] module at different scales and processed by Feature Pyramid Network [36] to generate features at different semantic levels. Similarly to other YOLO models, we adopt an anchor-based multi-scale detection scheme. Path Aggregation Network (PAN) [40] transfers positioning features bottom-up. We combine them with the features from FPN to obtain a better feature fusion effect and then directly use the multi-scale fusion feature maps in the PAN for detection. Thus, the detection heads predict bounding boxes and 3DMM parameters on different scales, ensuring high accuracy on different object sizes. This approach allows us to optimize FLAME parameters only on positive anchors, improving training efficiency. It also offers flexibility for applications that may not require full 3D meshes, allowing the extraction of bounding boxes in real time without computing full 3D mesh. The detection head in the YOLO-NAS model predicts the offset of the bounding box position and the scaling of the height and width, as well as the confidence of the prediction. We extend the detection head to also predict the 3DMM parameters by introducing six separate 3D parameter prediction modules, each consisting of two RepVGG blocks [6] and a final 1×1 convolution that predicts the final set of parameters. Each RepVGG block consists of three branches: a 3×3 Convolution followed by BatchNorm [25], a branch of a 1x1 Convolution with bias and a residual branch. Predicting different 3DDM parameters components separately achieves an extra level of disentanglement.

Table 9. **Model Architecture Comparison.** Analysis of different VGGHeads variants in terms of parameters, computational cost, and speed.

Model	Total Parameters	FLOPs (B)	FPS
VGGHeads-L	50,442,706	83.51	60.13
VGGHeads-M	32,378,236	52.01	69.38
VGGHeads-S	17,004,954	22.92	72.34

The weights of the loss components are set to $\alpha_1 = 50, \alpha_2 = 1, \alpha_3 = 1, \alpha_4 = 0.5, \alpha_5 = 2.5$ in the final version.

The framework is agnostic to the choice of backbone and can be adapted to use larger transformer models [42]. Yet, we observe the string performance on a smaller fully convolutional models and stick to this design to allow for real-time multi head mesh recovery. Furthermore the versatility of backbones allow to train even smaller models suitable for wide range of tasks and applications.

D. Head Pose Estimation

The full results on AFLW and BIWI datasets are presented in Tab. 10, Tab. 11. VGGHeads outperforms most of method optimized solely for 3D Head Pose Estimation even though it does not operate on tight head crops.

Figure 6. **Re-ID on Celeb-A.** We automatically detect and removed samples where face in the generated image is matched with one of the faces from Celeb-A dataset [41].

Figure 7. **Preserving Identity.** With subject neutral prompts diffusion Model blends feature of different people from the set it was trained on.

E. 3D Face Reconstruction

VGGHeads is more coarse than 3D reconstruction methods so detailed 3D face reconstruction is not the goal of our method and a standalone task in itself. Nonetheless, we evaluate the model on the Feng et al. benchmark [39]. The model achieves comparable results to other coarse 3D face reconstruction methods [19, 44, 55] despite not being optimized for shape and expression disentanglement. This performance is surprising since our method predicts the 3D face from a full image, not a tight crop as is typical for this task. While our method aims at solving a different task, it achieves good performance. Moreover, the VGGHeads dataset and model can be complementary to detailed face reconstruction methods such as DECA [17], MICA [80] or EMOCA [9], which often need to be initialized with a crop or initial coarse face shape.

F. Full Body Mesh Recovery

Our approach proves the viability of using synthetic data from diffusion models for body modeling, paving the way for future fully synthetic all-in-one methods. VGGHeads is the first step towards this goal. While full-body reconstruction methods [16, 72], achieved significant progress and attention, they often still rely on upstream face predictors and lack robustness in edge cases. To validate it we evaluate PIXIE's [16] performance on the BIWI [15] dataset for 3D Head Pose.

G. Controllable Generation

Conditioning the image generation on full head mesh helps to preserve the head shape and expression which is crucial for many AR applications. We trained the ControlNet [73] for SDXL [50] model that is conditioned on meshes recovered by our model. The meshes are rendered by mapping to RGB space with Projected Normalized Coordinate Code (PNCC)

Table 10. AFLW

Model	End to End	3DMM	MAE ↓	Pitch MAE ↓	Roll MAE ↓	Yaw MAE ↓
Dlib [29]	✗	✗	13.29	12.60	9.00	18.27
HopeNet [13]	✗	✗	6.16	6.56	5.44	6.47
6DRepNet [22]	✗	✗	3.61	4.58	2.98	3.27
RingNet [55]	✗	✓	8.27	4.39	13.51	6.92
3DDFA-V2 [19]	✗	✓	7.56	8.48	9.89	4.30
3DDFA [18]	✗	✓	7.39	8.53	7.39	5.40
DAD-3DHeads [44]	✗	✓	3.66	4.76	3.15	3.08
SynergyNet [63]	✗	✓	**3.35**	**4.09**	**2.55**	3.42
RetinaFace [12]	✓	✗	6.22	9.64	3.92	5.10
Img2Pose [3]	✓	✗	3.91	5.03	**3.28**	3.43
VGGHeads	✓	✓	**3.76**	**4.91**	3.37	**3.00**

Table 11. BIWI

Model	End to End	3DMM	MAE ↓	Pitch MAE ↓	Roll MAE ↓	Yaw MAE ↓
Dlib (68 points) [29]	✗	✗	12.25	13.80	6.19	16.76
HopeNet [13]	✗	✗	4.90	6.61	3.27	4.81
WHENet [77]	✗	✗	3.81	4.39	3.06	3.99
6DRepNet [22]	✗	✗	3.78	5.32	2.78	**3.23**
MNN [58]	✗	✗	**3.66**	4.61	**2.39**	3.98
3DDFA [18]	✗	✓	19.07	12.25	8.78	36.18
3DDFA-V2 [19]	✗	✓	8.81	12.08	7.54	6.80
RingNet [55]	✗	✓	7.34	5.37	7.82	8.82
DAD-3DNet [44]	✗	✓	3.98	5.24	2.92	3.79
RetinaFace [12]	✓	✗	4.49	6.42	2.97	4.07
Img2Pose [3]	✓	✗	**3.79**	**3.55**	3.24	4.57
VGGHeads	✓	✓	**3.79**	5.24	**2.65**	**3.47**

Model	3DRMSE↓	Median(mm)↓		Mean(mm)↓		Std(mm)↓	
		HQ	LQ	HQ	LQ	HQ	LQ
3DDFA-V2[19]	2.998	1.500	1.779	1.942	2.350	1.704	2.149
RingNet[55]	2.809	1.698	1.634	2.161	2.113	1.832	1.831
DAD-3DNet [44]	2.749	1.558	1.624	1.940	2.082	1.581	1.795
VGGHeads	2.996	1.622	1.801	2.079	2.353	1.801	2.054

Table 12. **Feng et al.[39]**

Table 13. **Comparison with Full Body Reconstruction [16].** VGGHeads achieves superior performance to full body recovery method [16] on BIWI 3D Head Pose Estimation. This shows that all in one methods still fail on challenging head understanding benchmarks.

Model	MAE↓	Pitch MAE↓	Roll MAE↓	Yaw MAE↓
PIXIE [16]	10.97	16.80	6.19	9.93
VGGHeads	**3.79**	**5.24**	**2.65**	**3.47**

[79], with the 3D coordinate of each vertex of the normalized head mesh encoded as RGB (NCCx = R, NCCy = G, NCCz = B). The heads on the generated images preserve the pose, expression and shape of the original photo Fig. 10.

H. Qualitative Results

Additional data samples from VGGHeads dataset are presented in Fig. 9.

We also include more visual results on AFLW [78] Fig. 12, BIWI [15] Fig. 13, DAD-3D [44] Fig. 14, WIDER [69] Fig. 8 and FDDB [26] benchmarks Fig. 11.

I. Limitations and Broader Impact

The dataset annotations are based on DAD-3D [44] so we don't aim to model neck, ears and eyeball vertices that are a part of the FLAME topology. The generation pipeline still can produce deformed small faces due to the limited resolution so we don't label and predict the 3D model parameters of the tiny faces. Also, the more advanced filtering methods and nsfw detection methods are a suitable venue for future explorations as on the large scale it is not feasible to guarantee the absolute correctness of the generated samples, even by adding the human evaluation into the process. By leveraging synthetic data generated through diffusion models, we reduce the privacy, ethics, and safety issues in human subject research, as no real personal data is used so that privacy and ethical standards are upheld. Furthermore, the synthetic dataset's high resolution and detailed annotations

Figure 8. **Qualitative Evaluation.** VGGHeads is able to accurately recover 3D head models on various complex scenes from WIDER Face [69] dataset.

provide a robust and versatile resource for developing and testing new models. This approach not only enhances the generalizability and accuracy of models trained on this data but also promotes ethical research practices by eliminating the need for real human subjects. The ability to generate large-scale synthetic datasets paves the way for safer and more inclusive research, free from the constraints and risks associated with real-world data collection. Thus, our work promotes ethical AI practices and sets a standard for future research in this area.

J. Fail Cases

The dataset fail cases are presented in Fig. 15. Typical failure cases are head detector failure (both FP and FN), misaligned mesh on out-of-distribution shapes and poses, severe occlusions, and deformed tiny faces in crowded scenes.

Figure 9. **Dataset Examples.** The synthetic data generation pipeline generates complex realistic real world scenes with multiple objects, covering wide range of poses and backgrounds while reducing age, gender and ethnical biases present in small real world datasets.

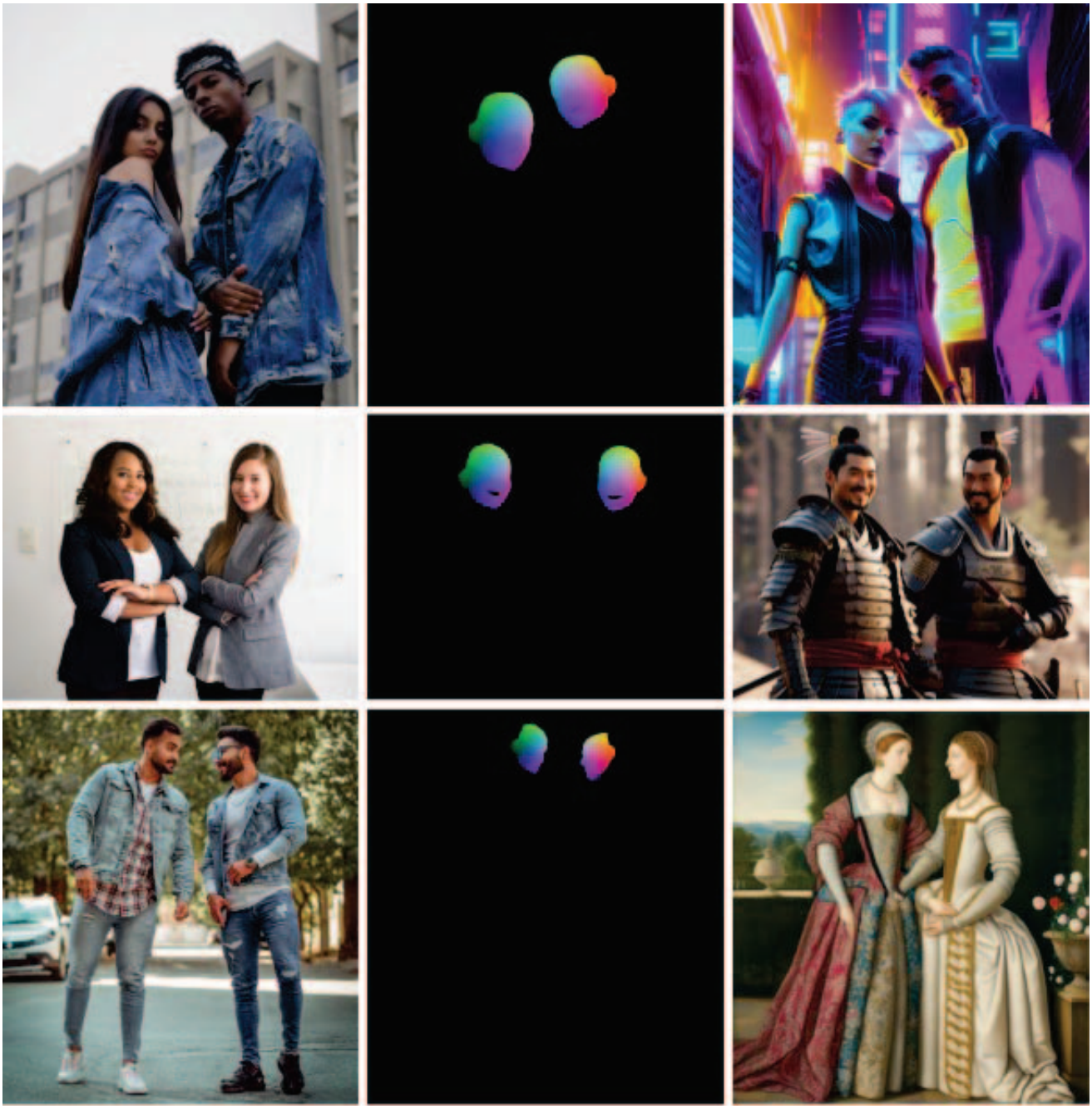

Figure 10. **ControlNet with VGGHeads.** The 3D condition provides a strong degree of control for the generative model, preserving shape, pose and expression of the input image.

Figure 11. **Qualitative Evaluation on FDDB.**

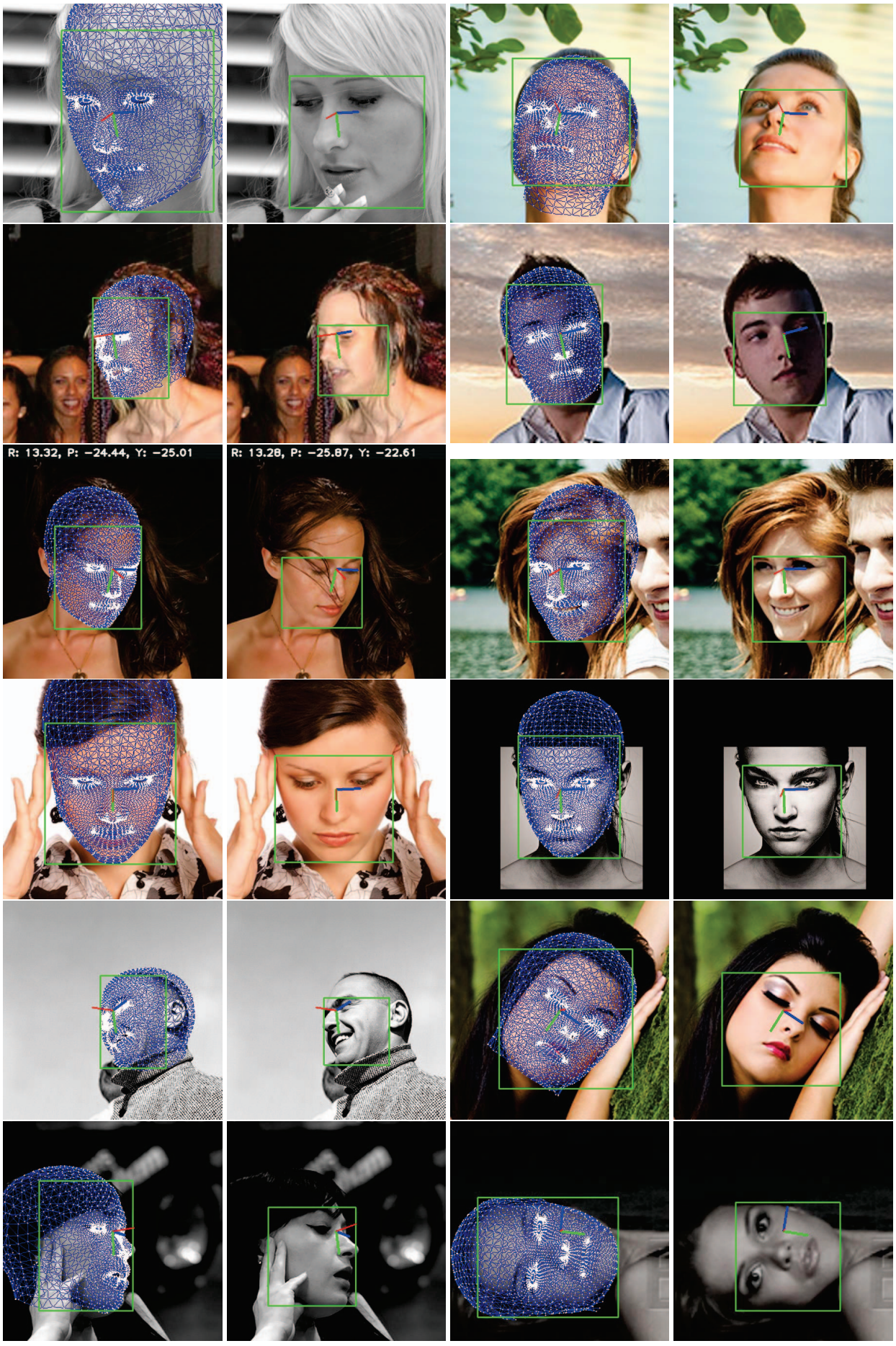

Figure 12. **Qualitative Evaluation on AFLW.**

Figure 13. **Qualitative Evaluation on BIWI.**

Figure 14. **Qualitative Evaluation on DAD-3D**

Figure 15. **VGGHeads Dataset Fail Cases**