Supplementary Material
Latent Space Imaging

Matheus Souza Yidan Zheng Kaizhang Kang Yogeshwar Nath Mishra
Qiang Fu Wolfgang Heidrich

KAUST

This supplementary document consists of three parts: I.) A more in-depth discussion of fundamental principles of Latent Space Imaging, including a discussion of the relationship to biological vision, and of alternative optical encoding schemes. II.) Additional implementation details of the current system, including the hardware prototype and training details. III.) Additional experimental results and ablation studies.

A Part I: Foundations of LSI

A.1 Relationship to Biological Vision

As mentioned in the main text, LSI is strongly inspired by biological vision, in particular the human visual system (HVS). Specifically, in the HVS, photoreceptors sense the spatial distribution of light, and retinal ganglion cells (RGCs) encode these spatial distributions into a compressed latent space transmitted over the optic nerve. Finally the visual cortex decodes and processes the information. The compression ratio between the number of photoreceptors and the number of axions in the optic nerve is about 100:1, and is enabled by the inherently non-linear processing of the RGCs.

LSI can be seen as analogous to this process: in our single pixel camera setup, the pixels of the spatial light modulator take on the role of the photoreceptors – they define the spatial resolution of the sensed image. The optical codes shown on the SLM, combined with the digital encoder take on the role of the RGCs, and both encode and compress this information, while the StyleGANXL takes on the role of the visual cortex.

The encoding module is therefore composed of a combination of a (linear) optical computing part and a non-linear digital encoder. The digital encoder is needed since optical computing [25] is in practice limited to linear operators. Critically, the most compressed representation is the intermediate latent space between the optical and the digital encoder, which enables highly compressed, low-bandwidth sampling hardware. We show that, with this two stage encoding, very high compression ratios can be achieved for the specific domain of facial images: ratios of 1:100 to 1:1000 for full image reconstruction, and even higher ratios for simpler tasks such as landmark detection or segmentation.

The principle of LSI can be applied to the latent space of any generative model, however the achievable compression ratios will depend on the complexity of the domain. To match the performance of the HVS on general vision tasks with a compression ration of 1:1000, we will need both a general AI generative model as well as a way to optically implement non-linear encodings.

A.2 Alternative Physical Implementations

In our prototype, we showcase the physical implementation using binary structural masking with a Digital Micromirror Device (DMD), where the light source is a monitor projecting the images. The SPI framework is an established platform for prototyping and testing different compressed imaging approaches, since the DMD acts as a programmable encoding element, making it easy to experimentally validate different code patterns and compression ratios.

However, SPI also has several downsides that limit its practicality for real-world imaging systems, including its form factor, and, most notably, its difficulty in dealing with moving scenes. However, the recent efforts in optical computing [25] have resulted in a range of alternative options for implementing the required linear optical encoding layer.

For example diffractive optical elements have been used to implement both convolutional [1] and fully connected linear layers [14], and meta-surfaces have been demonstrated for shift variant convolution kernels [24]. Due to the linear nature of light, all these methods are currently limited to linear operators. However, as we show with the SPI setup, linear optical encoders are sufficient for achieving high compression rates when combined with a small digital encoder. As such, any of the recently proposed optical compute frameworks could be used instead of the SPI framework, although typically this would require "freezing" the optical encoder into hardware. Nonetheless, for special purpose imaging systems, such an approach presents a practical avenue for developing real-world, snapshot-capable LSI systems.

In the longer term, LSI may be able to benefit from ongoing research efforts to develop non-linear optical computing hardware, for example based on non-linear optical materials, quantum effects, polarization, or other non-linear effects. Eventually, this may enable more powerful optical encoders, thereby reducing or even eliminating the need for the digital encoder module.

B Part II: Implementation Details

B.1 Digital Encoder Architecture

After acquiring the measurements *J* from the co-optimized optical encoder **O**, we employ a digital encoder network $\mathcal{D}_\theta$. The objective of this model is to match the intermediate latent space produced by the measurements with the actual latent able to reproduce the image of interest *I*.

The original StyleGANXL implementation receives a random noise vector and utilizes a mapping network to project this to different levels of details, result in $\mathbf{L} \in \mathbb{R}^{512\times l}$, l may vary depending the desired image resolution that the generative models was trained to generate. In the case of face images, l is equal to 18, finally we want **O** to map $J \in \mathbb{R}^d$ to $\mathbf{L} \in \mathbb{R}^{512\times 18}$. d corresponds to our measurement count, therefore, going from 512 to 4.

As depicted by Fig. 1, the intermediate latent goes first through 3 different set of blocks from left to right. The first one produces the first 3 elements of **L** out of 18, each one of them only take into account the initial *J*. The middle block, besides *J*, it also takes the concatenation of L_1, L_2 and L_3 through the MIX module, as input, outputting L_4 to L_7. Finally, the last block follows the same idea, it concatenates the output of the middle one and pass through another MIX block, the output of this is summed together with the original measurement and it passes through the block to finally result in L_8 to L_{18}. The idea here is to make the optical-aware inversion network match the multi-level structure of StyleGANXL, following a coarse-to-fine approach.

In the end, the latents from all detail levels pass through the MIX block. The purpose of the MIX blocks is to facilitate interactions and learn a weighted mixture across multiple detail levels. The model incorporates the spatial gating unit proposed in [15]. Within the MIX block (see Fig.1), the features are projected to a higher dimension, after which the resulting tensor x is split into two parts, u and v, along the channel dimension. This division allows for separate processing paths within the block. v is normalized and linearly projected using a weights matrix and bias vector to capture interactions in a static manner, as the weights matrix remains unchanged after training because it does not depend on the input. The final element-wise multiplication of u and the projected v modulates the information flow, controlling which parts of u are allowed to pass through—similar to a gating mechanism. Finally, this gated output is merged with the input through a residual connection, preserving the original information and enhancing gradient flow during backpropagation.

B.2 Loss Functions and Training Details

In this work, we trained the optical **O** and digital $\mathcal{D}_\theta$ encoders while keeping the generative model, StyleGANXL, frozen. To jointly optimize these components, we trained an off-the-shelf inversion network pre-trained to invert images to the latent space of a specific domain [19]. This network is used to compute a latent similarity loss function ($\mathcal{L}_{lat}$), making the output of the $\mathcal{D}_\theta$ similar to the latent space approximated by this model. This process is described in Equation 3 of the main text.

Additionally, we aim to match the image quality at the pixel level and enforce identity similarity. For this task, we used the identity loss ($\mathcal{L}_{id}$), which computes the cosine distance between feature maps extracted by ArcFace [6], a facial recognition network. We also utilized the ℓ_2 norm ($\mathcal{L}_{l2}$) and DINO/LPIPs features loss($\mathcal{L}_p$) [28, 31] to enforce pixel-wise and perceptual similarities, respectively. The details of DINO feature extraction are provided in [31].

The total loss function during training can be summarized as follows:

$$\mathcal{L}_{total} = \lambda_{lat}\mathcal{L}_{lat} + \lambda_{id}\mathcal{L}_{id} + \lambda_p\mathcal{L}_p + \lambda_{l2}\mathcal{L}_{l2} \quad (1)$$

After convergence, we added an additional loss term ($\lambda_{energy}\mathcal{L}_{energy}$) to account for the intensity diversity among the masks.

$$\mathcal{L}_{energy} = \tfrac{1}{d}\sum_j^d \left|\sum_i^{mn} \mathbf{O}_{i,j} - \epsilon_j\right| \quad (2)$$

Here, ϵ represents the ground-truth energy level, heuristically designed to correspond to a certain percentage of pixels set to one. This percentage increases incrementally from 10% to 90%, with a step of 1%. For strong compression from 1:1024 to 1:16384 the min and max energy boundaries move closer to 50% accordingly. The patterns are shuffled to avoid any undesired structure. As demonstrated in Fig. 2, with the energy loss applied, the pixel occupancy histogram showcases a broad spectrum of occupancies, reflecting a high diversity of patterns. In contrast, without this modification, the patterns are more uniform in their occupancy levels, exhibiting less variability.

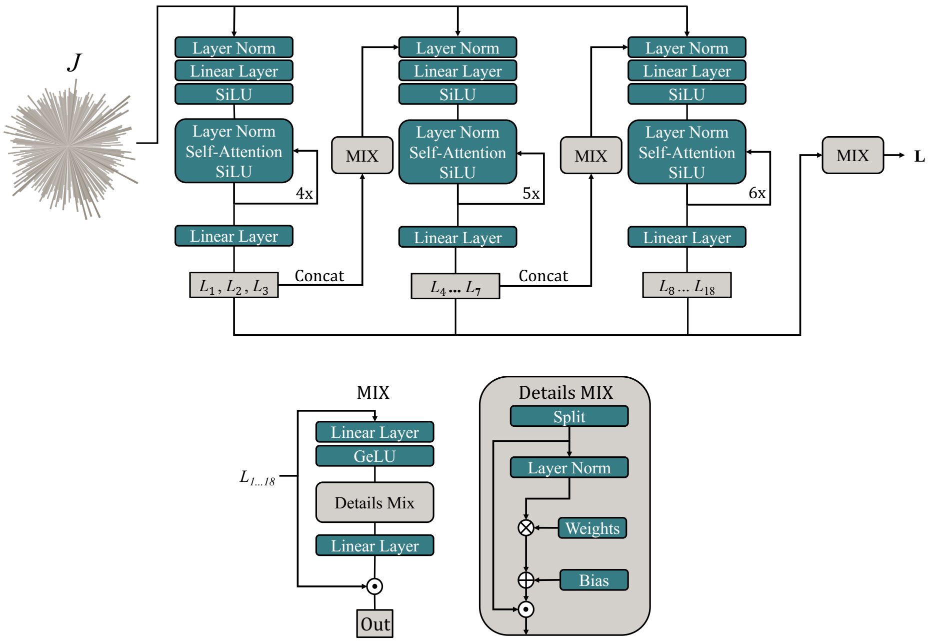

Figure 1. (Top) Overview of all the blocks involved in the digital encoder architecture. The three steps capture coarse, middle, and fine details, with each step depending on the output of the previous one. (Bottom) Illustration of the MIX process. The latent vector is statically projected to capture interactions among different levels of detail.

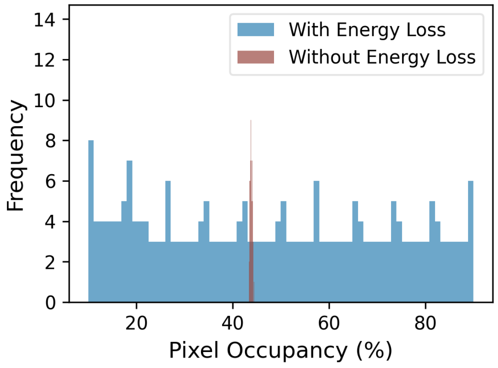

Figure 2. Shows the histogram of pixel occupancy, comparing scenarios where the energy loss function is and is not utilized. The use of energy modeling enhances the dynamic range of the measurements captured by the single sensor.

To optimize **O** and $\mathcal{D}_\theta$, we employ distinct optimizers: Lion [2] for **O** and Ranger (combination of Lookahead [27] and the Rectified Adam [16]) for $\mathcal{D}_\theta$, each with a learning rate of 10^{-4}. The weighting coefficients λ for each loss function are set at $1, 0.5, 0.8,$ and 1, in the respective order that they appear in Equation 1. Additionally, for the energy loss component, we set λ to 3. Regarding batch sizes, we typically use 32 for face datasets. However, for the AFHQ [3] dataset, as discussed in Sec. 3, the batch size is reduced to 8 due to the limited number of images available. For all experiments we utilize a single NVIDIA A100.

B.3 Optical Encoder Optimization Details

In the **O** optimization process, in order to deal with the quantized patterns, we employ the straight-through estimator (STE) [10], a technique also favored by VQ-VAEs [17, 18, 21]. During the forward pass, values are quantized and constrained before computing J. However, during backpropagation, gradients are allowed to flow through the non-quantized version, facilitating weight updates. Importantly, we ensure that the entries of **O** remain positive and are bounded between $[0, 1]$. Although using complementary patterns [26] allows for the inclusion of negative values, empirical experiments have shown that this approach does not yield significant improvements and necessitates double the measurements.

Initially, masks are generated from a uniform distribution and are then binarized through a quantization process. This process results in binary patterns that achieve an even total counting of 0s and 1s, theoretically maximizing bit entropy and steering the system towards an optimal solution [13]. This distribution presumes the sensor is sensitive enough to discern very fine differences, a demanding prerequisite as each pattern will have similar total intensity. This issue was addressed using the energy loss mechanism.

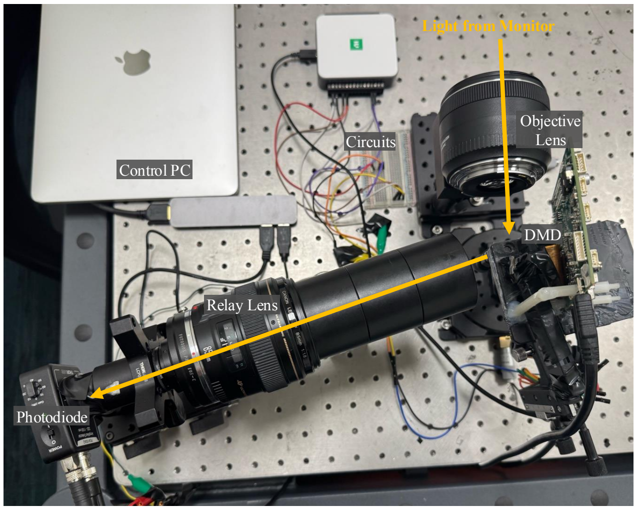

Figure 3. Experiment setup. An objective lens collects light from monitor and relays it to the DMD, which modulates the light with reflecting patterns. A relay lens directs the modulated light to the photodiode, and the PC records the measurements. Yellow arrows denote the light path.

B.4 Experimental setup

Our experimental setup utilizes a single-pixel imaging configuration as follows: Face images are displayed on a high-dynamic-range Eizo CG3145 monitor. These images are then focused onto the DMD of the Texas Instruments DLPLCR4500EVM evaluation board using a Canon EF 35mm f/2 IS USM objective lens. We interpolate our learned 256×256 patterns to optimally fill the largest possible area, adjusting the scale to compensate for the diamond-shaped micromirror grid of the DMD. This adjustment ensures full utilization of the height within a central square region. The DMD spatially modulates the image by reflecting patterns according to an optimized encoding matrix ($\mathbf{O}$). The resultant patterned light is captured by a photodiode (PDA100A2) through a Canon EF 85mm f/1.8 USM relay lens. An analog-to-digital converter (NI 6001) converts the received analog signals into digital form. The digital data are then received by a connected PC, which is also responsible for synchronizing all components involved in the experimental setup.

To enhance the congruence between our measurements and simulations, we employed a white image to determine a global scaling deviation and applied a correction factor. This approach effectively reduces the gap between simulation and measurement, yet some inherent challenges persist due to non-linear behaviors, minor deviations in DMD reflection angles, and sensor noise, which complicate the precise replication of simulated conditions. However, leveraging the controlled environment of our experimental setup, we are able to systematically gather real-world training data and fine-tune $\mathcal{D}_\theta$.

The fine-tuning process involves selecting a subset of 200 training images and measuring their outputs using our configured setup. We employ the loss functions specified in Equation 1, while reducing the learning rate to 10^{-5}, to optimize performance and accuracy in real-world applications.

C Part III: Additional Results

We present additional LSI results encompassing a diverse set of images, along with the range of compression levels.

Figures 7, 8, and 9 demonstrate a consistent retention of facial details, particularly as compression increases, highlighting the effectiveness of our approach at higher compression rates. By targeting the latent space, our method addresses the over-smoothing often observed with aggressive sub-sampling strategies, preserving textures and producing realistic reconstructions instead of flat, featureless outputs. Notably, key facial features such as eyebrow shape, beards, and smiles are well reconstructed, emphasizing the persistence of facial expressions. In contrast, competitor models fail to achieve similar results.

For the downstream tasks, Fig. 6 presents additional results from the **experimental setup** for attribute classification. To further evaluate and enhance the assessment of our method, we include an extensive set of qualitative comparisons (see Figs. 10, 11, 12, 13, 14, and 15) against other methods. These comparisons focus on simulated results for face segmentation and landmark detection.

Table 1 quantitatively summarizes our results, highlighting the overall superiority of our method in reconstruction and downstream tasks, particularly in high-compression setups.

Additional domains were explored in Fig. 4, we utilize our simulated pipeline with 512 (1:128) measurements to reconstruct cats and dogs images from AFHQ dataset [3]. These datasets imposed additional challenge because they are very small compared with FFHQ, with only 5000 training images. However, LSI is capable to faithfully reconstruct such domain pictures with a correspondingly trained encoder.

C.1 Out-of-Distribution Cases and Limitations

As our focus is on faces, and our underlying generative model is primarily trained on frontal-centered faces under typical daily lighting conditions, significant deviations from these conditions are not well-modeled and often result in noticeable hallucinations. Fig. 5 illustrates several out-of-domain scenarios. For instance, providing a non-facial input, such as a cat, to the face generative model results in a

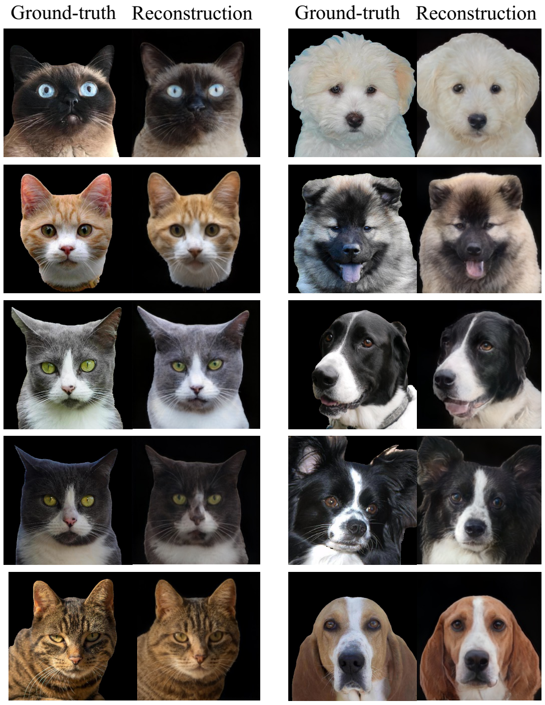

Figure 4. Illustrates another set of reconstructions from our simulated results, highlighting the versatility of our methods across various domains, such as cats and dogs. For this dataset we utilized StyleGAN2 instead XL, also showing the versatility of our method utilizing different generative models.

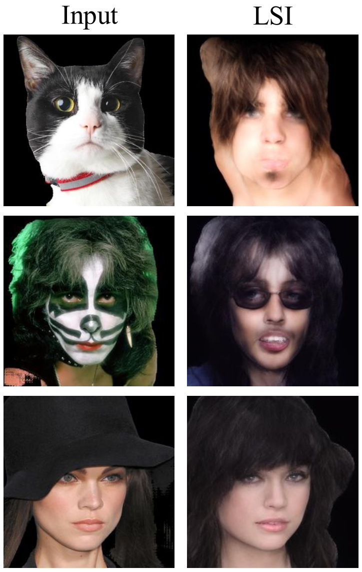

Figure 5. Illustrates the conditions where LSI fails to reproduce the scene since the target scene is out of distribution of the generative model.

hallucinated face that aligns with the characteristics of the given input signal. Similarly, accessories like hats or heavy makeup cannot be accurately modeled, as such data lies outside the model's training distribution.

C.2 Qualitative and Quantitative Comparisons

We evaluated a diverse set of representative models encompassing various approaches, including block-based sensing matrices, unfolding networks, single-pixel imaging techniques, and deep learning-driven single-pixel frameworks. Additionally, we incorporated optimization-based methods that leverage generative models as priors. To provide a broader perspective, we introduced a variant of our pipeline, referred to as **non-latent**, which replaces the generative model and latent representation with a state-of-the-art reconstruction network and a conventional signal recovery method. Further details on this model are provided below.

Preliminary notes on comparisons:

- All methods were retrained on the same dataset and evaluated on the same testing dataset as LSI.
- SAUNet, FSI-DL, and Non-latent LSI utilize a quantized (binary) sensing matrix, similar to LSI. In contrast, the sensing matrices of other methods can represent any float value, giving them an (unrealistic!) advantage.
- SAUNet, OCTUF [20], and TCS-Net [8] are trained to reproduce luminance only, as proposed in their original papers. During inference, chrominance components were added using ground-truth values. Other methods, however, consider color images, modeling them either implicitly (via latent space) or explicitly (via demosaicking).
- For comparison purposes, downstream applications for all competitors were evaluated using FaRL [30], based on their reconstructed images, as they lack the inherent capability to perform these tasks. Because of low reconstruction quality, some competitor methods may fail the downstream tasks, we discard such data points only for the specific method. Notably, for LSI, these results were achieved through a simple linear projection, entirely eliminating the need for a separate, complex model tailored to each specific task.

AuSamNet and FSI-DL [9] Both are Fourier basis methods with deep learning reconstruction algorithms; the first optimizes the mask, and the second utilizes a fixed heuristically designed circular mask. The patterns P_ϕ are generated using the ideal proposed by [29]

$$P_\phi(x, y; f_x, f_y) = a + b\cos(2\pi f_x x + 2\pi f_y y + \phi), \quad (3)$$

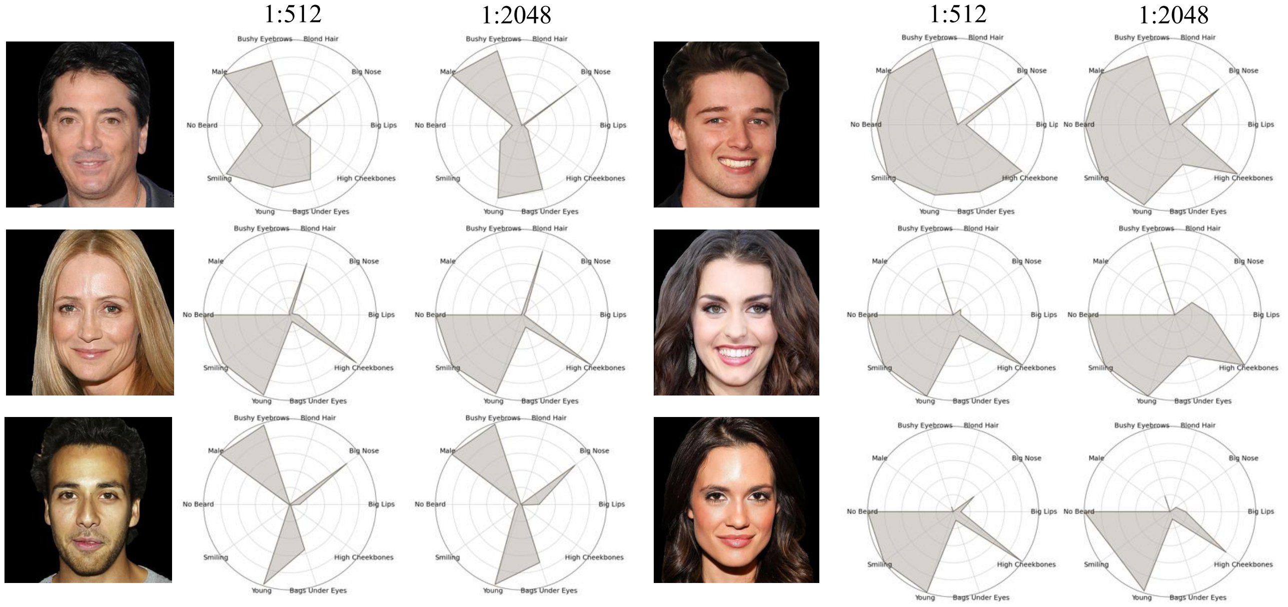

Figure 6. Additional attribute visualizations from the experimental validation are presented. The leftmost column displays the original faces as shown on the monitor, followed by reconstructed images at compression ratios of 1:512 and 1:2048, which correspond to 128 and 32 measurements, respectively.

where (x, y) represents the 2D Cartesian coordinates in the scene, a symbolizes the average intensity distribution, b stands for the amplitude of the Fourier basis pattern, and (f_x, f_y) indicates the non-zero spatial frequency points within the optimized mask. Furthermore, ϕ denotes the initial phase, adopting three steps phase shifting of 0, $2\pi/3$, and $4\pi/3$.

This work also utilizes a color filter array (CFA) and demosaic process to reduce the number of measurements necessary to retrieve color images, computing the Fourier transform of the CFA and multiplying it with the mask.

When evaluating under our heavily compressed setting, AuSamNet cannot optimize the mask, leading to unstable results. FSI-DL is capable of reconstructing the images but produces low-quality results as demonstrated by Figures 7, 8, 9.

SAUNet [23] This approach proposes a 2D measurement system and an unfolding network as a reconstruction algorithm. Their measurement can be defined by the equation $Y = HXW^T$, where H and W are learned during the optimization process, and X is the input. To enable a fair comparison, we clamped and then quantize their measurement matrix during training to be binary, following the same procedure adopted by our methods. We add a normalization layer after utilizing H and W to avoid exploding gradients and instability.

OCTUF [20] and TCS-Net [8] Both methods utilize transformer-based reconstruction techniques integrated with block-based image compressed sensing for measurements, where the image is divided into smaller patches. However, the sensing matrices used are not restricted to values below 1, making their physical implementation unfeasible. The methods were trained using the original code without any modifications, apart from adjustments to the compression ratios. Finally, they benefit from an unbounded sensing matrix and similar to SAUNet [23], they do not model color, only learning the luminance.

Optimization-Based (CSGM [11]) leverages a generative model as a prior to reconstruct compressed signals. However, its measurement matrix is neither bounded nor quantized, making it unsuitable for physical deployment. Additionally, it requires multiple iterations and cannot achieve reconstruction with a single forward pass.

Non-Latent LSI We further evaluate our approach by implementing a version of LSI that does not utilize the latent space. In this configuration, the 2D image $\in \mathbb{R}^{256\times256}$ is reconstructed directly from the measurements using Differential Ghost Imaging (DGI) [7], followed by a state-of-the-art

Method	Comp. Ratio	VGGFace↑	DLib↑	FID↓	Acc.↑	$F1$↑	NME_{dg}↓
LSI	1:128	91.97%	92.74%	**27.38**	**89.07%**	70.00%	1.48
	1:256	**90.98%**	**92.68%**	**26.62**	**89.15%**	70.94%	1.43
	1:512	**89.61%**	**91.67%**	**28.66**	**89.20%**	**70.25%**	**1.48**
	1:1024	**81.12%**	**87.44%**	**28.79**	**88.74%**	**69.18%**	**1.52**
SAUNet	1:128	39.47%	64.12%	58.96	83.05%	74.15%	1.36
	1:256	17.84%	50.12%	75.53	82.66%	71.32%	1.59
	1:512	5.52%	31.42%	104.23	81.08%	66.79%	1.99
	1:1024	2.86%	16.59%	107.35	79.98%	62.37%	2.20
FSI-DL	1:128	15.28%	32.60%	107.48	81.66%	62.20%	2.00
	1:256	3.30%	13.16%	118.79	79.43%	53.38%	3.20
	1:512	1.82%	1.93%	134.40	76.72%	41.45%	5.06
	1:1024	0.85%	2.05%	173.71	75.96%	34.95%	6.06
OCTUF	1:128	54.99%	63.54%	70.83	83.93%	75.78%	1.31
	1:256	59.27%	73.09%	77.16	83.99%	**74.78%**	**1.26**
	1:512	16.63%	43.31%	96.17	82.33%	69.92%	1.73
	1:1024	5.83%	26.62%	99.85	81.13%	66.18%	1.99
TCS-Net	1:128	25.15%	49.69%	119.42	82.21%	66.00%	1.71
	1:256	1.32%	5.50%	248.60	77.71%	52.11%	3.31
	1:512	1.07%	2.62%	296.48	75.19%	N/A	7.43
	1:1024	0.11%	1.62%	333.34	74.05%	N/A	15.07
CSGM	1:128	20.04%	51.85%	44.46	82.17%	62.20%	1.89
	1:256	13.66%	39.16%	46.30	81.28%	56.96%	2.36
	1:512	5.87%	21.96%	53.12	78.68%	46.07%	3.29
	1:1024	2.00%	10.23%	63.70	76.16%	38.31%	4.15
Non-Latent	1:128	**94.22%**	**94.84%**	72.96	84.47%	**77.47%**	**1.07**
	1:256	87.52%	91.49%	75.62	83.71%	74.59%	1.26
	1:512	71.82%	82.98%	81.87	83.08%	70.10%	1.53
	1:1024	41.47%	68.22%	89.04	82.09%	64.73%	1.90

Table 1. Latent Space Imaging (LSI) and competing methods were quantitatively evaluated on simulated results across various downstream tasks, with compression ratios ranging from 1:128 to 1:1024. The results emphasize the superior performance of LSI, particularly in highly compressed scenarios. For the image reconstruction task, evaluations were conducted using the VGGFace [5] and DLib [12] image recognition pipelines. Metrics assessed include Fréchet Inception Distance (FID), classification accuracy, F1-mean score for segmentation, and Normalized Mean Error (NME), normalized by the diagonal of the face bounding box. Bold numbers highlight the best model for each metric and setting.

image reconstruction model [4], similar to the methodology described by [22]. This replaces the latent space representation and generative model entirely with a direct reconstruction pipeline. The training process, including perceptual and identity loss functions, Optical Encoder (**O**) optimization, and other parameters, remain unchanged. While this version performs well for lower compression ratios (1:128), it is consistently outperformed by LSI, particularly at higher compression levels such as 1:512 and 1:1024 (see Table 1), underscoring the critical role of latent space representation.

	VGGFace↑	Dlib↑	FID↓
LSI Random	48.9%	71.7%	37.20
LSI	**90.98%**	**92.68%**	**26.62**

Table 2. Quantitative comparison between LSI and LSI with fixed random patterns (akin to classical Compressed Sensing), emphasizing the critical role of jointly optimizing the optical encoder for enhancing the performance. Compression ratio of 1:256.

C.3 Background Discussion

To emphasize facial features, we mask out the background to prevent our already highly compressed signal from being used on non-facial information. It is worth noting, however, that our method remains effective even when trained on faces with uncontrolled backgrounds, achieving identity classification accuracies of 86.26% and 86.88% with VGGFace and DLib, respectively. Ablation study conducted for a compression ratio of 1:256.

C.4 Optical Encoding vs. Fixed Random Encoding

Optimizing the optical encoding plays a critical role in image reconstruction. This is evident in Table 2, which shows a significant performance drop when LSI is trained using fixed random patterns instead optimized encoding towards the latent space.

References

[1] Julie Chang, Vincent Sitzmann, Xiong Dun, Wolfgang Heidrich, and Gordon Wetzstein. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.*, 8(1):12324, 2018. 1

[2] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3, 4

[4] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1239–1248, 2022. 7

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4685–4694, 2019. 7

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4690–4699, 2019. 2

[7] Fabio Ferri, D Magatti, LA Lugiato, and A Gatti. Differential ghost imaging. *Physical review letters*, 104(25):253603, 2010. 6

[8] Hongping Gan, Minghe Shen, Yi Hua, Chunyan Ma, and Tao Zhang. From patch to pixel: A transformer-based hierarchical framework for compressive image sensing. *IEEE Transactions on Computational Imaging*, 9:133–146, 2023. 5, 6

[9] Wenxin Huang, Fei Wang, Xiangyu Zhang, Ying Jin, and Guohai Situ. Learning-based adaptive under-sampling for fourier single-pixel imaging. *Opt. Lett.*, 48(11):2985–2988, 2023. 5

[10] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. *Adv. Neural Inform. Process. Syst.*, 29, 2016. 3

[11] Ajil Jalal, Liu Liu, Alexandros G Dimakis, and Constantine Caramanis. Robust compressed sensing using generative models. *Advances in Neural Information Processing Systems*, 33:713–727, 2020. 6

[12] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 7

[13] Mingbao Lin, Rongrong Ji, Zihan Xu, Baochang Zhang, Fei Chao, Chia-Wen Lin, and Ling Shao. Siman: Sign-to-magnitude network binarization, 2022. 3

[14] Xing Lin, Yair Rivenson, Nezih T Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406):1004–1008, 2018. 1

[15] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to MLPs. *Adv. Neural Inform. Process. Syst.*, 34: 9204–9215, 2021. 2

[16] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Int. Conf. Learn. Represent.*, 2019. 3

[17] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[18] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019. 3

[19] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[20] Jiechong Song, Chong Mou, Shiqi Wang, Siwei Ma, and Jian Zhang. Optimization-inspired cross-attention transformer for compressive sensing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6174–6184, 2023. 5, 6

[21] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. 3

[22] Fei Wang, Chenglong Wang, Chenjin Deng, Shensheng Han, and Guohai Situ. Single-pixel imaging using physics enhanced deep learning. *Photon. Res.*, 10(1):104–110, 2022. 7

[23] Ping Wang and Xin Yuan. Saunet: Spatial-attention unfolding network for image compressive sensing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5099–5108, 2023. 6

[24] Kaixuan Wei, Xiao Li, Johannes Froech, Praneeth Chakravarthula, James Whitehead, Ethan Tseng, Arka Majumdar, and Felix Heide. Spatially varying nanophotonic neural networks. *Sci. Adv.*, 10(45), 2024. 1

[25] Gordon Wetzstein, Aydogan Ozcan, Sylvain Gigan, Shanhui Fan, Dirk Englund, Marin Soljačić, Cornelia Denz, David AB Miller, and Demetri Psaltis. Inference in artificial intelligence with deep optics and photonics. *Nature*, 588 (7836):39–47, 2020. 1

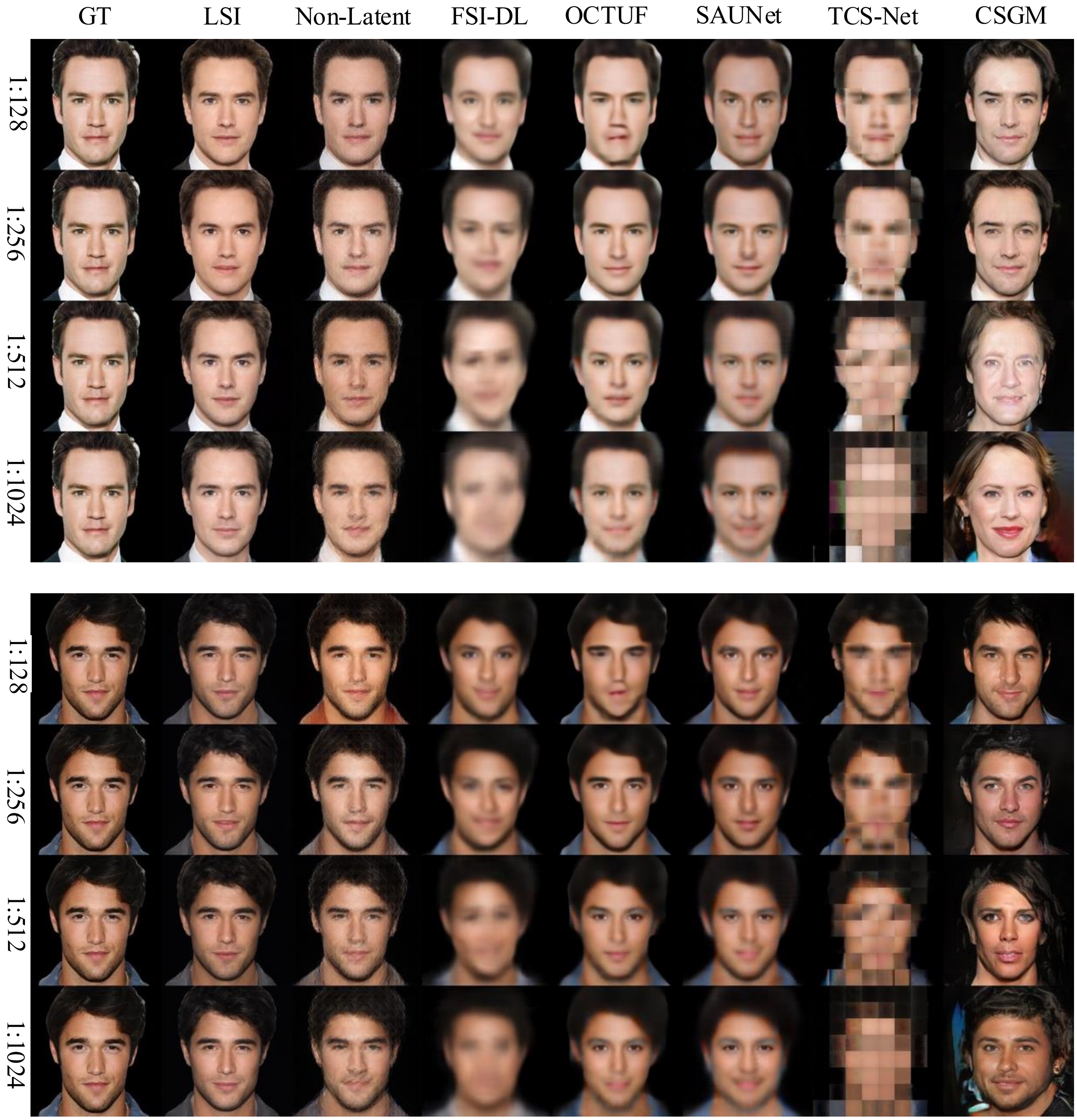

Figure 7. Different methods compared for facial reconstruction quality.

[26] Yibo Xu, Liyang Lu, Vishwanath Saragadam, and Kevin F Kelly. A compressive hyperspectral video imaging system using a single-pixel detector. *Nat. Commun.*, 15(1):1456, 2024. 3

[27] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Adv. Neural Inform. Process. Syst.* Curran Associates, Inc., 2019. 3

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 2

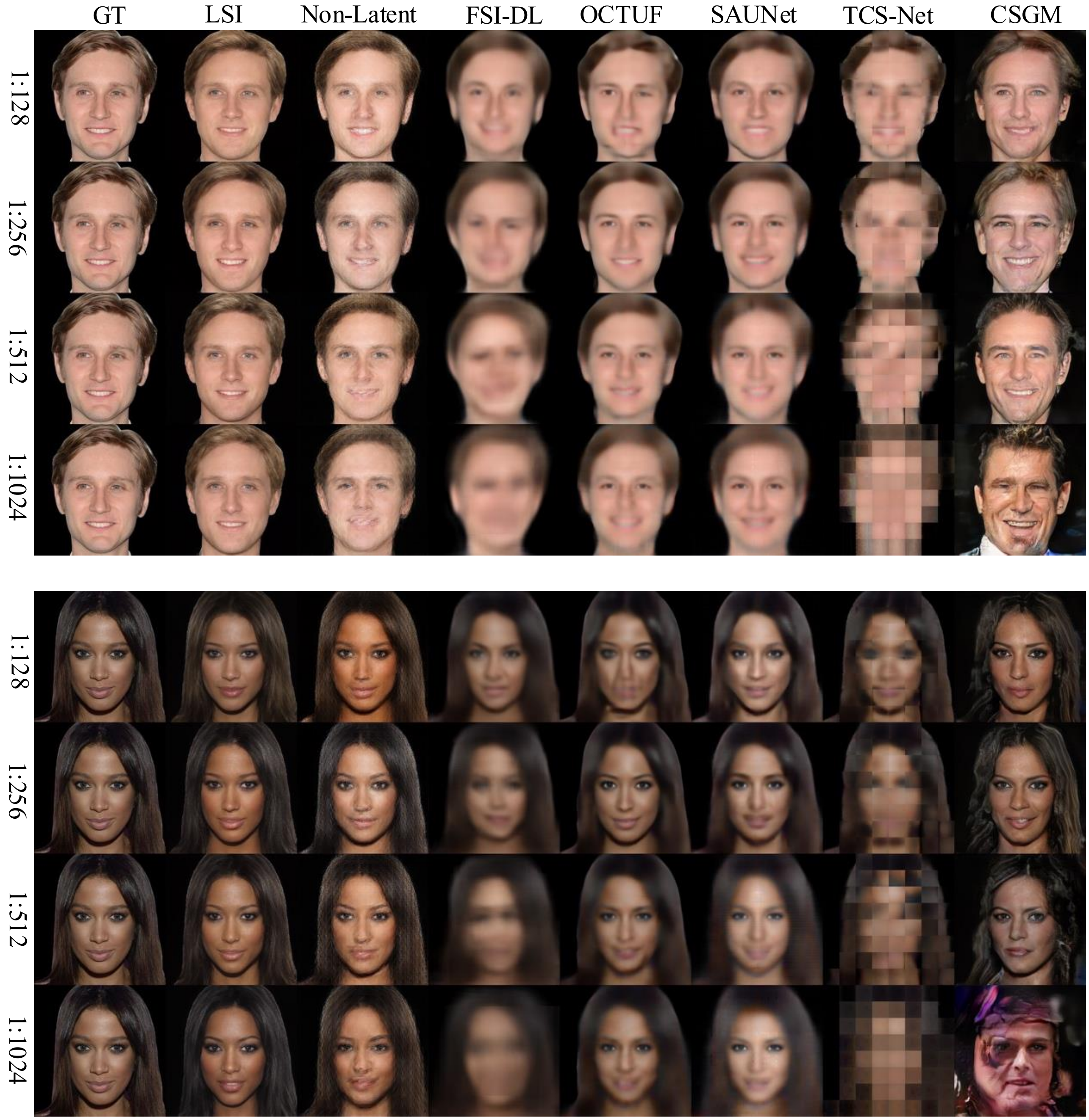

Figure 8. Different methods compared for facial reconstruction quality.

[29] Zibang Zhang, Xiao Ma, and Jingang Zhong. Single-pixel imaging by means of fourier spectrum acquisition. *Nat. Commun.*, 6(1):6225, 2015. 5

[30] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709, 2022. 5

[31] Yang Zhou, Zichong Chen, and Hui Huang. Deformable one-shot face stylization via dino semantic guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*

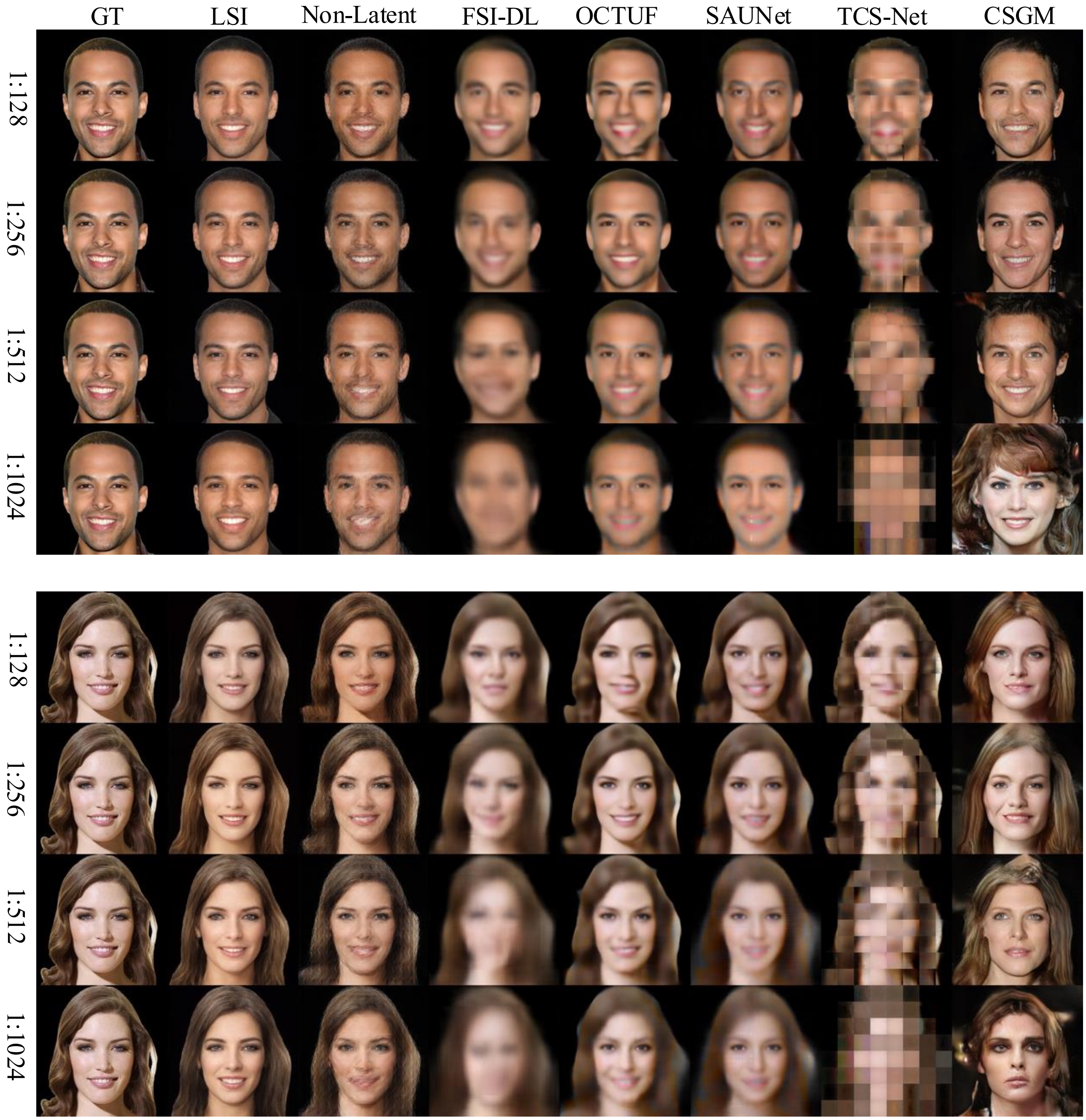

Figure 9. Different methods compared for facial reconstruction quality.

Pattern Recognition, pages 7787–7796, 2024. 2

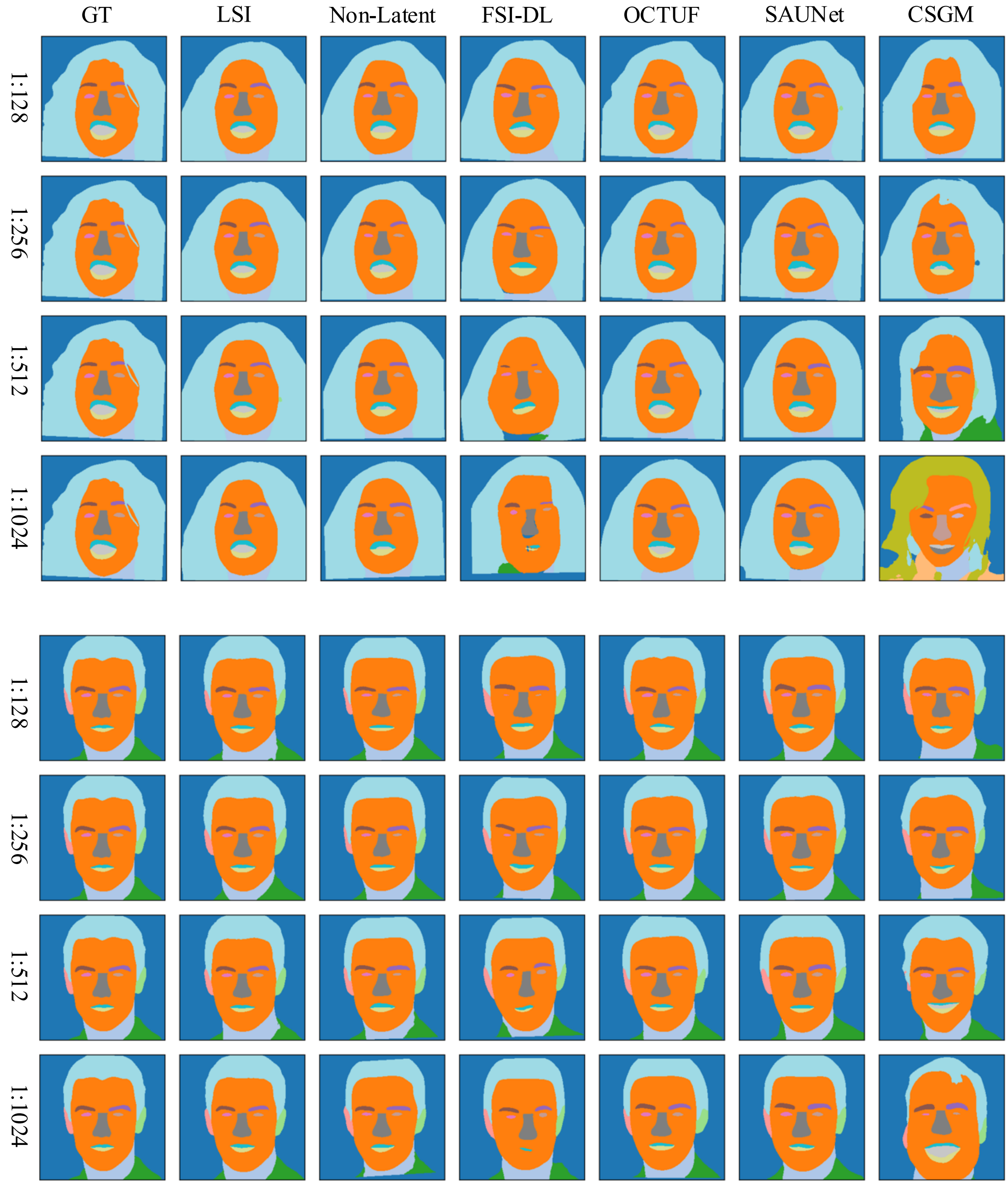

Figure 10. Facial segmentation comparison among different methods.

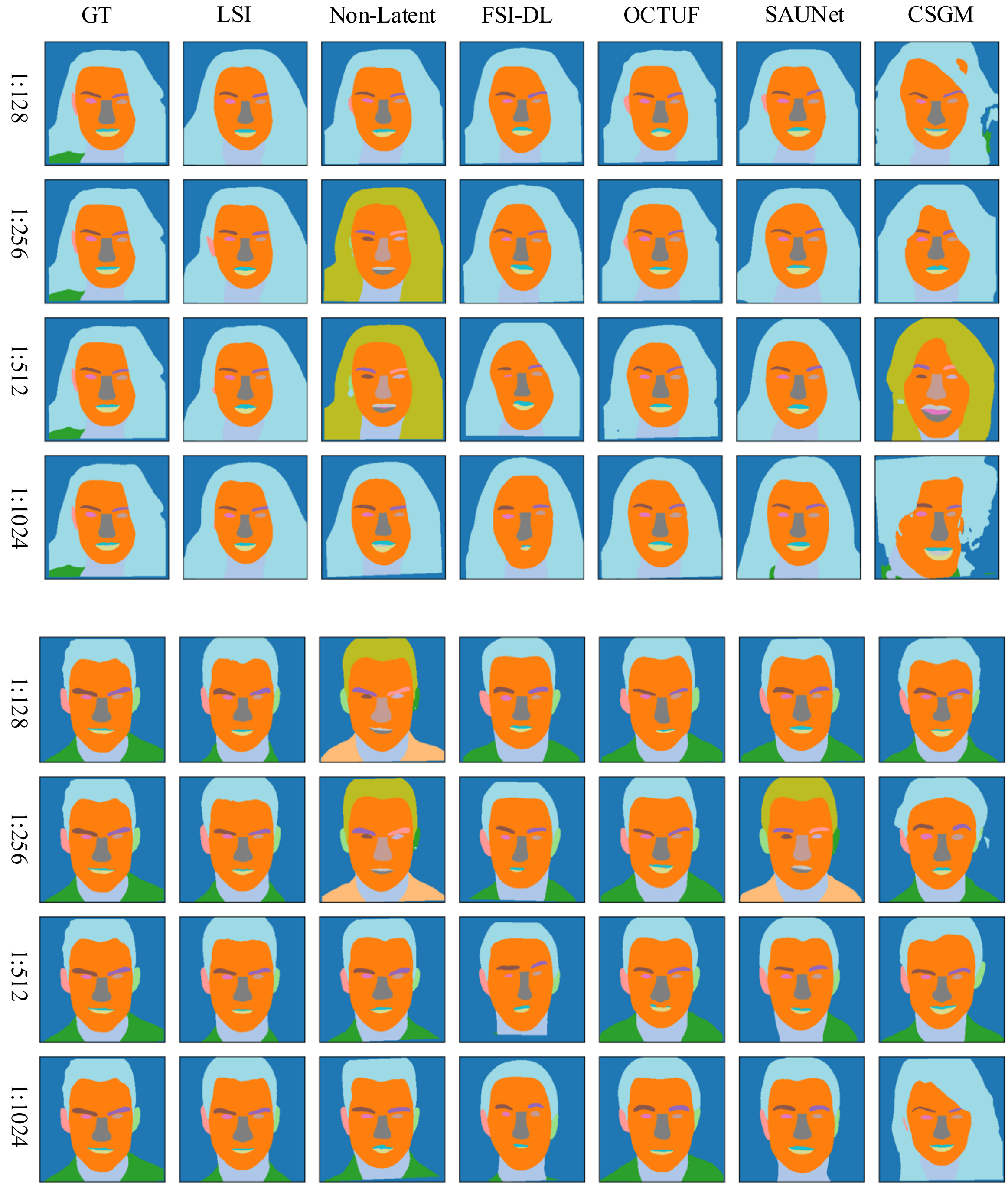

Figure 11. Facial segmentation comparison among different methods.

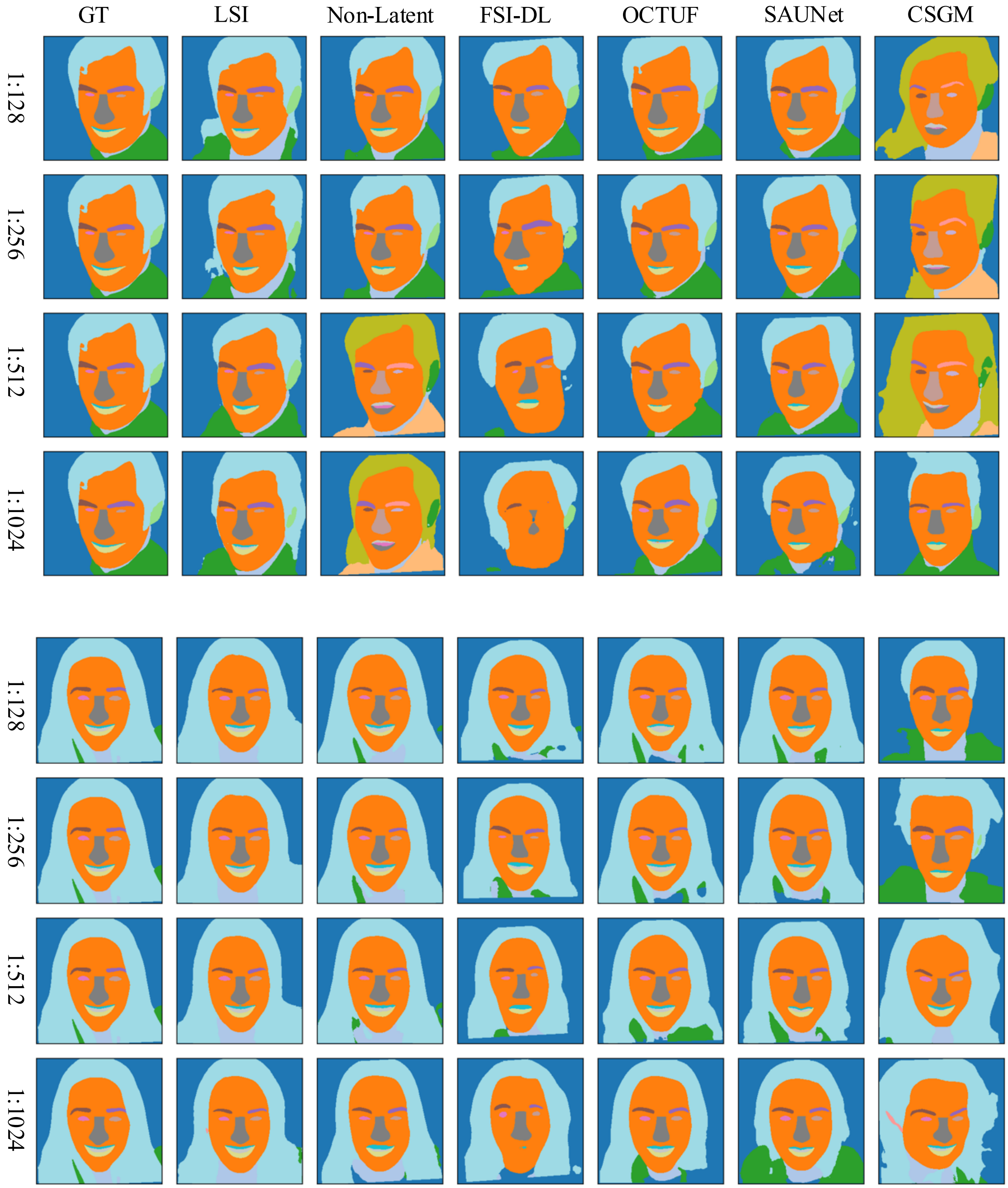

Figure 12. Facial segmentation comparison among different methods.

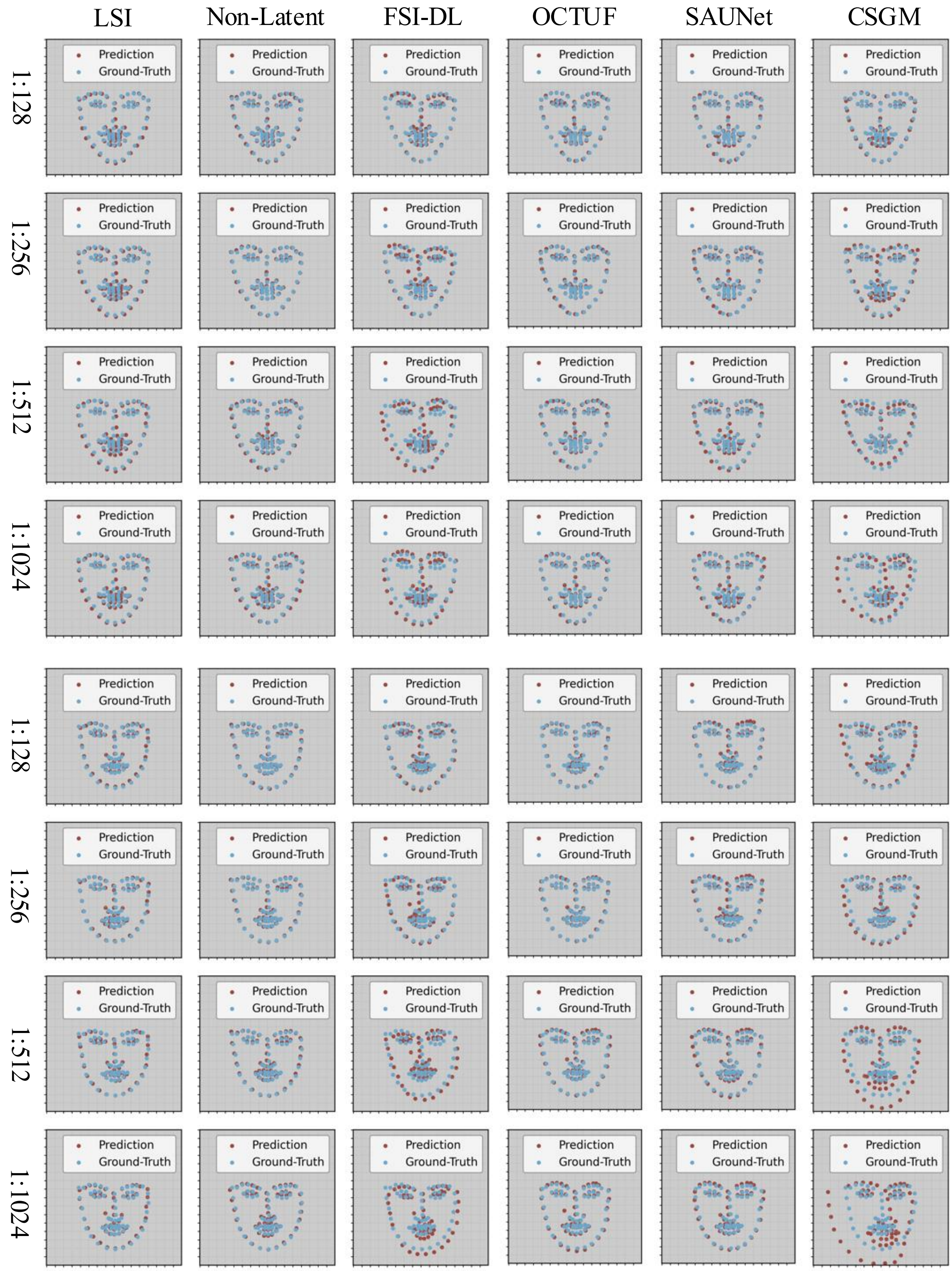

Figure 13. Landmarks detection comparison among different methods.

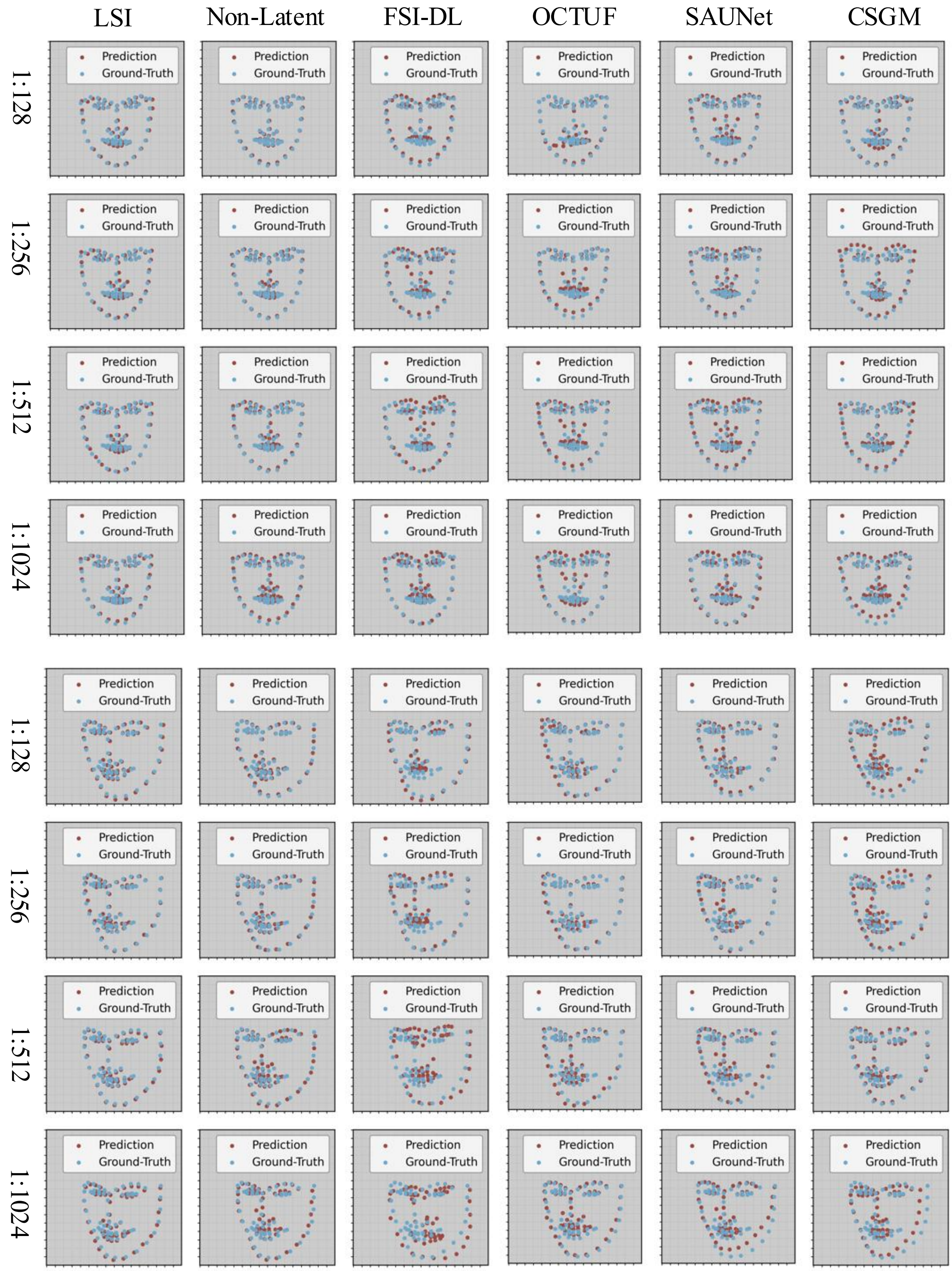

Figure 14. Landmarks detection comparison among different methods.

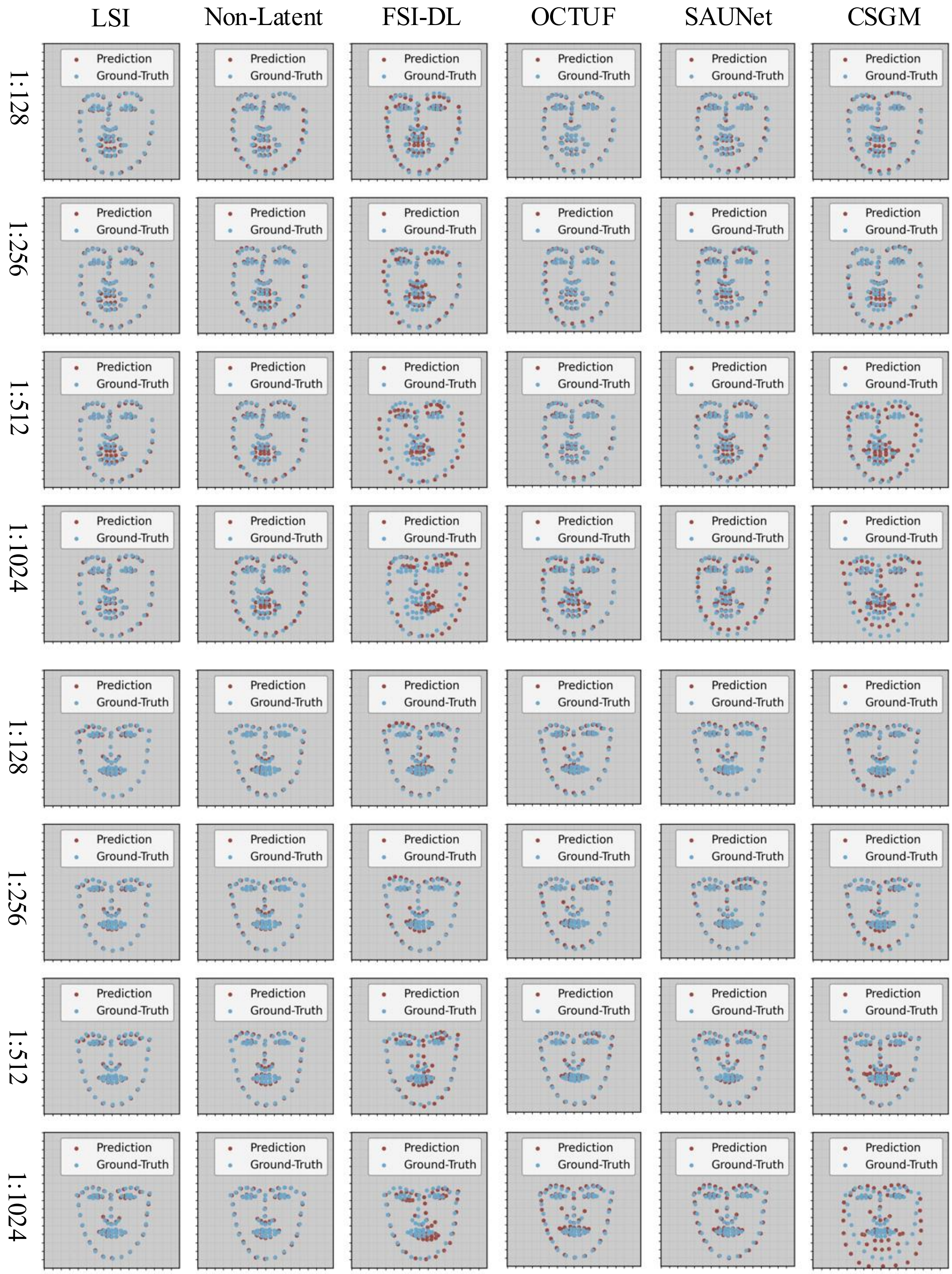

Figure 15. Landmarks detection comparison among different methods.