Supplementary Materials of Auto-CARD: Efficient and Robust Codec Avatar Driving for Real-time Mobile Telepresence

Yonggan Fu[1], Yuecheng Li[2], Chenghui Li[2], Jason Saragih[2], Peizhao Zhang[2], Xiaoliang Dai[2], Yingyan (Celine) Lin[1]
[1]Georgia Institute of Technology [2]Meta
{yfu314, celine.lin}@gatech.edu {yuecheng.li, leo.li, jsaragih, stzpz, xiaoliangdai}@meta.com

Table 1. Benchmark the searched encoders with SOTA encoder designs in terms of measured latency on Quest 2/Pixel 3 and rendering MSE across different 10 identities and 3 view directions.

Iden.	Model	EEM	EEM -ch50	EEM -res50	AVE-L (Ours)	AVE-M (Ours)	AVE-S (Ours)
	MFLOPs	2930.77	765.38	747.44	605.14	306.93	**174.75**
	Lat. (ms) Quest 2	12.48	10.02	9.40	4.59	3.26	**2.47**
	Lat. (ms) Pixel 3	483.47	164.27	117.27	70.53	52.61	**37.78**
S1	Front	8.48	8.54	11.27	**6.91**	7.46	7.54
	Left	8.11	8.36	10.84	**6.80**	7.29	7.41
	Right	8.04	8.07	10.65	**6.45**	7.01	7.03
S2	Front	15.70	16.08	22.77	**14.63**	15.10	16.21
	Left	14.52	15.08	21.08	**13.74**	14.56	15.22
	Right	17.51	18.00	24.82	**16.17**	16.31	17.53
S3	Front	12.03	12.85	15.53	**10.91**	11.42	12.25
	Left	12.00	12.88	14.96	**10.93**	11.48	11.99
	Right	12.73	13.62	16.45	**11.41**	11.83	13.05
S4	Front	17.42	18.71	21.40	**15.62**	16.33	16.98
	Left	19.12	20.41	23.22	**16.80**	17.57	18.01
	Right	17.47	18.72	21.27	**15.56**	16.08	16.94
S5	Front	7.01	7.81	15.95	**5.78**	5.94	6.06
	Left	7.32	8.05	16.09	**6.22**	6.33	6.45
	Right	7.10	7.93	15.45	**5.92**	6.03	6.24
S6	Front	19.52	20.77	25.05	**17.34**	18.39	19.08
	Left	26.28	26.47	33.76	**22.27**	24.33	24.34
	Right	15.89	16.73	22.25	**14.27**	15.41	15.55
S7	Front	19.52	20.77	25.05	**17.34**	19.08	18.39
	Left	10.03	10.23	17.00	**9.00**	9.39	9.43
	Right	10.14	10.16	18.26	**9.02**	9.31	9.54
S8	Front	8.44	12.32	18.32	**6.22**	6.62	7.61
	Left	8.94	12.98	18.57	7.19	**7.03**	8.12
	Right	8.39	12.15	18.46	**6.55**	6.61	7.56
S9	Front	4.12	4.72	5.97	**3.46**	3.75	3.96
	Left	4.46	4.61	5.82	**3.71**	4.03	4.00
	Right	6.63	7.09	9.06	**5.31**	5.83	6.15
S10	Front	10.52	11.97	12.75	**9.37**	9.79	9.94
	Left	10.38	11.75	12.40	**9.25**	9.75	9.89
	Right	11.01	12.57	12.75	**9.89**	10.18	10.33
Avg.	-	11.96	13.01	17.24	**10.47**	11.01	11.43

1. More Quantitative Rendering Results

Rendering MSE across all identities. In addition to the 6 identities evaluated in Sec. 5.2 of the main text, we further validate our AVE-NAS searched encoders across another 4 identities and summarize all the rendering MSE in Tab. 1. We can observe that our searched encoders consistently achieve better MSE-efficiency trade-offs across identities as compared to the SOTA EEM encoder [3] based on the measurement on Meta Quest 2, e.g., our searched AVE-L achieves a 2.72× speed-up over EEM while also reducing the rendering MSE by 1.49 on average and our AVE-S achieves a 5.05× speed-up with an average MSE reduction of 0.53.

Rendering LPIPS and FID: We also measure the rendering quality in terms of LPIPS [6] and FID [2] (the lower the better), which better aligns with human perception. We benchmark our AVE-L model with the baseline EEM model across six identities. As shown in Tab. 2, our AVE-L model still outperforms the baseline EEM model, e.g., a 0.39 FID reduction in average with 2.72× speed-up on Quest 2, indicating that our method can consistently achieve better rendering quality in terms of human perception.

Table 2. Benchmark our AVE-L model with the baseline EEM model across six identities in terms of the achieved LPIPS and FID.

Identity	Front View				Left View			
	LPIPS ↓		FID ↓		LPIPS ↓		FID ↓	
	Gabe	Auto-CARD	Gabe	Auto-CARD	Gabe	Auto-CARD	Gabe	Auto-CARD
S1	0.0740	**0.0457**	4.121	**2.506**	0.0503	**0.0459**	3.382	**2.821**
S2	0.0707	**0.0694**	2.969	**2.954**	0.0713	**0.0701**	3.010	**2.864**
S3	0.0711	**0.0667**	2.260	**2.107**	0.0693	**0.0650**	1.857	**1.805**
S4	0.0669	**0.0651**	2.772	**2.730**	0.0666	**0.0630**	2.570	**2.367**
S5	0.0709	**0.0604**	5.061	**4.355**	0.0754	**0.0651**	3.524	**3.119**
S6	0.0707	**0.0615**	3.437	**2.930**	0.0666	**0.0620**	2.869	**2.595**

2. Detailed Experiment Setup

Training settings. We follow the same setting in [4] to train the decoders, which is fixed during encoder search and training. For encoder search and training, we adopt the same training schedules for architecture parameters and model weights. In particular, we search/train the encoder for 50K/100K steps, respectively, using an Adam optimizer with a batch size of 16 and an initial learning rate of 1e-3 decayed by 0.1 every 40K steps. The weighted coefficients for $\mathcal{L}_{latent}$, $\mathcal{L}_{gaze}$, $\mathcal{L}_{geo}$,$\mathcal{L}_{tex}$, $\mathcal{L}_{kpt}$, and $\mathcal{L}_{ren}$ in Eq. (4)

Table 3. Visualize the operators, channel scales, and input resolutions of searched encoders (zoom-in for better view).

Model	Input Resolution	View	Searchable Factor	Feature Extraction Backbone	Branch			MFLOPs (Total)
					Latent Code	Gaze	Key Point	
AVE-L	80	Left Eye	Operator Channel Scale MFLOPs	['fuse-mb', 'conv'] [0.53125, 0.5] 46.08	['fuse-mb', 'skip', 'conv', 'conv', 'fuse-mb', 'fuse-mb'] [0.53125, 0.5, 0.53125, 0.5, 0.75, 0.75] 43.67	['conv', 'fuse-mb', 'conv', 'fuse-mb', 'conv', 'conv'] [1.0, 0.5, 0.5, 0.5, 0.9375, 0.9375] 35.12	['fuse-mb', 'skip'] [0.5, 0.5] 54.85	605.14
		Right Eye	Operator Channel Scale MFLOPs	['fuse-mb', 'conv'] [0.8125, 0.875] 101.12	['conv', 'skip', 'conv', 'skip', 'conv', 'fuse-mb'] [0.5, 0.5, 0.5, 0.5, 0.75, 1.0] 69.92	['conv', 'conv', 'conv', 'conv', 'conv', 'fuse-mb'] [1.0, 0.5, 0.6875, 0.5, 0.75, 0.9375] 34.13	['fuse-mb', 'skip'] [0.5, 0.5] 54.85	
		Mouth	Operator Channel Scale MFLOPs	['fuse-mb', 'fuse-mb'] [0.53125, 0.5] 34.77	['conv', 'conv', 'conv', 'fuse-mb', 'fuse-mb', 'fuse-mb'] [0.5, 0.5, 0.53125, 0.53125, 1.0, 1.0] 129.43	- - -	- - -	
AVE-M	64	Left Eye	Operator Channel Scale MFLOPs	['fuse-mb', 'conv'] [0.53125, 0.5] 38.88	['fuse-mb', 'skip', 'skip', 'skip', 'conv', 'conv'] [0.5, 0.5, 0.5, 0.5, 0.75, 0.75] 31.35	['fuse-mb', 'conv', 'conv', 'fuse-mb', 'fuse-mb', 'conv'] [1.0, 0.5, 0.5, 0.5, 1.0, 1.0] 24.84	['skip', 'skip'] [0.5, 0.5] 5.31	306.93
		Right Eye	Operator Channel Scale MFLOPs	['conv', 'conv'] [0.5, 0.5] 37.67	['skip', 'skip', 'conv', 'skip', 'conv', 'conv'] [0.5, 0.5, 0.5, 0.5, 1.0, 1.0] 57.96	['conv', 'conv', 'conv', 'fuse-mb', 'fuse-mb', 'fuse-mb'] [0.5, 0.5, 0.5, 0.5, 0.875, 0.9375] 16.09	['skip', 'skip'] [0.5, 0.5] 5.31	
		Mouth	Operator Channel Scale MFLOPs	['conv', 'skip'] [0.5, 0.5] 28.07	['conv', 'skip', 'conv', 'conv', 'conv', 'fuse-mb'] [0.5, 0.5, 0.53125, 0.5, 1.0, 1.0] 60.25	- - -	- - -	
AVE-S	64	Left Eye	Operator Channel Scale MFLOPs	['conv', 'skip'] [0.5, 0.53125] 28.07	['skip', 'skip', 'skip', 'skip', 'fuse-mb', 'conv'] [0.5, 0.5, 0.5, 0.5, 0.75, 0.75] 10.91	['fuse-mb', 'skip', 'conv', 'fuse-mb', 'conv', 'conv'] [0.5625, 0.5, 0.53125, 0.5, 0.75, 0.9375] 6.14	['skip', 'skip'] [0.5, 0.53125] 5.31	174.75
		Right Eye	Operator Channel Scale MFLOPs	['conv', 'skip'] [0.53125, 0.53125] 28.07	['skip', 'skip', 'skip', 'skip', 'conv', 'fuse-mb'] [0.5, 0.5, 0.5, 0.5, 0.75, 1.0] 14.65	['fuse-mb', 'skip', 'conv', 'conv', 'conv', 'fuse-mb'] [0.5625, 0.5, 0.5, 0.5, 0.75, 0.9375] 6.82	['skip', 'skip'] [0.5, 0.53125] 5.31	
		Mouth	Operator Channel Scale MFLOPs	['conv', 'skip'] [0.5, 0.5] 28.07	['fuse-mb', 'skip', 'conv', 'conv', 'conv', 'fuse-mb'] [0.5, 0.5, 0.5, 0.5, 0.75, 1.0] 40.21	- - -	- - -	

of the main text are set to 1e-1, 1, 1, 1, 1e3, and 1e-4, respectively, for balancing their magnitudes. The training of the whole Codec Avatar pipeline [4] takes about 12 GPU days and the search and training of avatar encoders take 2 GPU days.

Hyper-parameters of our proposed Auto-CARD. Setup of AVE-NAS: For the sampling frequency K for resolution search in Sec. 4.2.3, we set it as 16, which can help stabilize the resolution search process based on our empirical experiments. For the search objective of AVE-NAS, we set τ and m in Eq. (3) of the main text, which are the temperature parameter and the momentum factor, respectively, as 10 and 0.9 across all the experiments. Setup of LATEX: For LATEX introduced in Sec. 4.3.2 of the main text, we extrapolate the latent codes based on the previous 4 frames (T=4). To avoid error accumulation across frames caused by inaccurate extrapolation, we also set a rule that if the latent codes of the previous 3 frames are derived by LATEX, we enforce an encoder inference for acquiring the latent code of the current frame. In addition, LATEX's lightweight prediction head, featuring one convolutional layer, one pooling layer, and one fully-connected layer, is applied after the convolutional head in the feature extraction backbone (introduced in Sec. 3) to early predict the latent code, which is used to adaptively decide whether to perform extrapolation via comparing with a threshold determined by the target skip ratio.

3. Visualize the Encoder Architectures

SOTA EEM encoder. The SOTA EEM encoder [4], which requires 2.9 GFLOPs to encoder each frame, adopts a feature extractor backbone to extract the features of HMC captured images, on top of which a latent code branch, a gaze branch, and a key point branch, are applied to estimate the latent code, gaze, and key points of the current frame, respectively. In particular, the input resolution is [192, 192, 3] which concatenates the captured infrared images of the left eye/right eye/mouth views. The feature extractor backbone is composed of one convolution head and two bottleneck blocks in [1] with an output channel number of [64, 128, 128], respectively, and a total stride of 4; The latent code branch comprises 6 bottleneck blocks featuring a total stride of 8 and an output channel number of [128, 128, 256, 256, 256, 256], respectively, and one fully-connected layer to map the latent features into a 256-d vector; The gaze branch comprises 6 bottleneck blocks featuring a total stride of 8 and an output channel number of [128, 128, 128, 128, 64, 64], respectively, plus one pooling layer and a fully-connected layer to map the gaze features to a 6-d vector (i.e., 3-d gaze directions for each eye); The key point branch comprises 4 bottleneck blocks with an output channel number of [128, 128, 64, 64], respectively, followed by a convolutional layer to regress 19 key points.

Search space of our AVE-NAS. To ensure sufficient flexibility of the encoder architecture, the search space spans operator types, depth, width, and input resolution. For each view in the view-decoupled supernet, we set 2, 6, 6, 2 searchable blocks for the feature extractor backbone, latent code branch, gaze branch, and key point branch with a maximal output channel of [64,64], [64, 64, 256, 256, 512,512], [256, 256, 128, 128, 32, 32], and [128, 64], respectively. For the supported operator types, each searchable block can be set as Fused-MBConv [5], a single convolution with a kernel size of 3×3, or skip connections. Note that if there exists a feature dimension mismatch, the skip connection is switched to a convolutional layer with a kernel size of 1×1. The encoder depth is searchable via setting the operators to skip connections and the searchable encoder width is achieved

by applying an output channel scale, which is searched from [0.5, 0.53125, 0.5625, 0.59375, 0.625, 0.6875, 0.75, 0.8125, 0.875, 0.9375, 1.0], on top of the maximal channel numbers. For input resolution search, the candidates are [32, 48, 64, 80, 96, 128, 192] where the maximally allowed resolution is inherited from EEM [4].

Our searched encoders. We visualize the searched encoders, including AVE-L, AVE-M, and AVE-S, in Tab. 3. We can observe that the searched branches for different views feature diverse complexity, aligning with the fact that different cameras can capture different aspects of facial appearance and motion. This indicates that the view-decoupled supernet design not only echos the future distributed encoding trend, but also enables the flexibility of view-specific encoder structures and thus improves the achievable accuracy-efficiency trade-off of avatar encoding.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[3] Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. The eyes have it: An integrated eye and face model for photorealistic facial animation. *ACM Transactions on Graphics (TOG)*, 39(4):91–1, 2020. 1

[4] Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. The eyes have it: An integrated eye and face model for photorealistic facial animation. *ACM Trans. Graph.*, 39(4), 2020. 1, 2, 3

[5] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021. 2

[6] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1