Supplemental Material: Expressive Telepresence via Modular Codec Avatars

Hang Chu[1,2] Shugao Ma[3] Fernando De la Torre[3]
Sanja Fidler[1,2] Yaser Sheikh[3]

[1]University of Toronto [2]Vector Institute [3]Facebook Reality Lab

1 Section 4.2 Further Illustrations

In this supplemental section, we provide more illustration for Section 4.2 in our main paper on Exemplar-based Latent Alignment. Fig. 1 (a) shows a diagram showing our alignment process. Dome-captured data are reconstructed and tracked to obtain 3D faces, while correspondences between headset-captured images and avatar are established via [1]. Dome-captured 3D faces are masked to train modular VAEs, which produces the modular decoder $\mathcal{D}_k^{\text{mask}}$ and exemplar codes $\mathbf{C}_k^{\text{mask}}$. Finally, $\hat{\mathbf{c}}_k^{\text{part}}$ is obtained via exemplar-based matching as described in Eq.(5) of our main paper.

The necessity of exemplar-based latent alignment roots from a fundamental challenge in VR telepresence: the face is occluded by the VR headset. Because of this challenge, it is physically infeasible to obtain unobstructed full face observations with the headset. This makes correspondence establishing approach [1] the best available alternative to approximate animation ground-truth. However, significant domain and content gap exist between dome-captured faces and headset-captured faces, e.g., differences in lighting condition, facial skin pressure from wearing the headset, appearance change between capture runs such as sweat and facial hairs, etc. This means to obtain $\hat{\mathbf{c}}_k^{\text{part}}$, pure image-based synthesis results in large discrepancy in the latent space, as shown in Fig. 1(b). Furthermore, when multiple capture runs exist for the same person, the discrepancy becomes more severe as different runs have further domain and content differences.

The usage of exemplar-based latent alignment effectively locks $\hat{\mathbf{c}}_k^{\text{part}}$ to the fixed bases defined by dome-captured codes $\mathbf{C}_k^{\text{part}}$. This essentially achieves a trade-off between lowering the image-based synthesis residual error, and avoiding final codes that are irregular or spurious.

2 Dataset Further Details

We provide further details for the headset captured sessions in Table 1 that are used in our experiments. In total there are 14 sessions captured for 4 persons (named as **p1**, **p2**, **p3**, **p4**) with three headsets in 6 different environment (named as **office**, **control**, **monitor**, **desk**, **usual**, **flash**). Each check mark represents one capture session and for the testing sessions the check marks are shaded. It

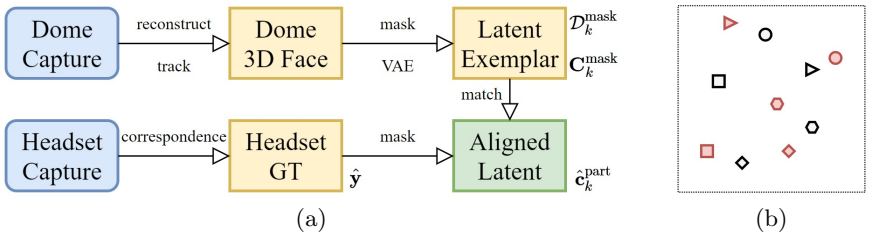

Fig. 1. Further illustrations for Sec.4.2 Exemplar-based Latent Alignment. (a) shows the steps of our described alignment process to obtain $\hat{\mathbf{c}}_k^{\text{part}}$. (b) shows the latent space comparing exemplars $\mathbf{C}_k^{\text{part}}$ (black) and headset code obtained through pure image-based synthesis without using exemplars (red), symbols represent different performed expressions. Pure image-based synthesis often lead to irregular and spurious codes.

is worth noting that the capture environments for testing sessions never appear in the training sessions. For person **p1** and **p2**, the headsets used in testing are also never used in training. These details further illustrate the robustness of our method against different capture headsets and/or capture environments.

3 Ablation Study Further Details

In this supplemental section, we provide further details and discussions for the ablation study shown in Table 2 of our main paper.

- **blend** For the ablation version we set the blending weights equal among all modules across the face. Using blending weights improves the result significantly by 0.66, showing the importance of blending modules adaptively according to the dynamic expression.
- **end2end** For the ablation version we train the encoders and synthesizers separately using intermediate supervision without end-to-end fine-tuning. End-to-end training improves the result significantly by 0.57, showing the importance of treating MCA model as a whole and enabling interaction between its sub-networks.
- **soft-ex.** For the ablation version, we replace $\hat{\mathbf{c}}_k^{\text{part}}$ with one-hot categorical vectors and train the encoders with cross-entropy loss. This has a significant impact on performance by 0.39, showing the effectiveness of using part codes obtained by our exemplar-based latent alignment.
- **dimen.** For the ablation version, we use a latent dimension of 16 instead of 256. This affects the performance by 0.07, as it is difficult for lower dimensional codes to capture subtle expression differences.
- **skip-mod.** For the ablation version, we remove the cross-module connections between the modular branches of image encoders. This affects the performance by 0.08, showing the usefulness of exploiting the correlations between modular images.

	office				control				monitor				desk				usual				flash			
person	p1	p2	p3	p4	p1	p2	p3	p4	p1	p2	p3	p4	p1	p2	p3	p4	p1	p2	p3	p4	p1	p2	p3	p4
headset1	✓					✓*																		
headset2			✓	✓	✓	✓					✓	✓		✓					✓	✓				
headset3																					✓	✓		

Table 1. Further details for different headset capture runs. Shaded cell denotes testing capture runs. Starred cell denotes two capture runs with different facial hair density.

– **tconv.** For the ablation version, we replace the two temporal convolution layers with two single-frame fully connected layers. This affects the performance by 0.04, showing the usefulness of exploiting temporal correlations between frames.

4 Video Demonstration

Please refer to the attached supplemental video for more qualitative examples of the MCA model, as well as the two extensive applications Eye Amplification and Flexible Animation as described in Section 5.3 of our main paper.

5 Limitations and Future Work

Besides failure cases of our model as shown in Fig. 7 in our main paper, we list a list of limitations and potential improvements in order to facilitate future work. We are particularly thankful to ECCV reviewers for their insightful suggestions and discussions.

– In our experiments, we have empirically set the modules to correspond each single camera. Other ways to separate the modules, such as having both eyes to share the same module, are worth investigating.
– Our current model is person-specific, as is the standard approach for hyper-realistic avatars. Our dataset is a step forward by including multiple rooms, captures, and devices, but still limited in the number of subjects due to the increased data collection cost per subject. The generalization of codec avatar methods is an important direction for our future work.
– We focus on the talking and conversation videos in our evaluation because they possess more variation across multiple capture runs. For the videos where the subject performs fixed facial expressions, we found MCA does not significantly outperform CA. This is due to these expressions have high similarity between training and testing capture runs.
– In eye amplification, we simply applying a multiplier on the difference between current modular latent code and the base latent code corresponding to the closed eye. This sometimes causes unwanted subtle gaze movement as shown in 2:45 to 2:50 of our video. This shows the latent space is non-linear to the actual physical motion, which suggests a potential solution to further rectify the latent space so that it better reflects the dynamics of facial parts.

- Our current $\mathbf{v}_0$ is set as a single frontal-facing view. It can be potentially useful to use other view or multiple views, which provides further information such as the side contours of the jaw.
- Our ground-truth expression is obtained via using more cameras on the tracking headset along with iterative optimization based on photometric errors. This is the best approximation to the true expression and we find this provides sufficiently accurate results. Better ground-truth acquisition with more convenient hardware setup can be further investigated in the future.

References

1. Wei, S.E., Saragih, J., Simon, T., Harley, A.W., Lombardi, S., Perdoch, M., Hypes, A., Wang, D., Badino, H., Sheikh, Y.: Vr facial animation via multiview image translation. In: SIGGRAPH. (2019)