FusionVAE: A Deep Hierarchical Variational Autoencoder for RGB Image Fusion (Supplementary Material)

Fabian Duffhauss[1,2], Ngo Anh Vien[1], Hanna Ziesche[1], and Gerhard Neumann[3]

[1] Bosch Center for Artificial Intelligence
{fabian.duffhauss,anhvien.ngo,hanna.ziesche}@bosch.com
[2] University of Tübingen
[3] Karlsruhe Institute of Technology
gerhard.neumann@kit.edu

A Implementation details

For FusionMNIST and FusionT-LESS, we model the decoder's output by pixel-wise independent Bernoulli distributions. For FusionCelebA, we use pixel-wise independent discretized logistic mixture distributions as proposed by Salimans et al. [73].

The residual cells of the encoder are composed of batch normalization layers [69], Swish activation functions [72], convolutional layers, and Squeeze-and-Excitation (SE) blocks [68] as proposed in [74]. In the decoder, we also follow [74] and build the residual cells out of batch normalization layers, 1x1 convolutions, Swish activations, depthwise separable convolutions [67], and SE blocks. However, we omitted normalizing flow because in our experiments it showed to increase the training time without improving the prediction accuracy significantly.

For each dataset, we chose the size of the architecture individually to achieve acceptable accuracy while keeping the training time reasonable. Tab. 6 provides details about the used hyperparameters.

Hyperparameter	FusionMNIST	FusionCelebA	FusionT-LESS
# latent groups per scale	5, 2	10, 5, 2	10, 5, 2
spatial dimensions of $\boldsymbol{z}_l$ per scale	$4^2, 8^2$	$8^2, 16^2, 32^2$	$8^2, 16^2, 32^2$
# channels in $\boldsymbol{z}_l$	10	20	20
# GPUs	2	4	2
# training epochs	400	90	500
batch size	800	32	32
Training time	4h	48h	28h

Table 6: Main hyperparameters of our experiments.

In general, the number of latent groups L should be chosen depending on the complexity of the task at hand. We made our decision based on the L of

the NVAE [74] but reduced it for computational reasons. For FusionCelebA and FusionT-LESS, we use 17 latent groups, for FusionMNIST only seven. Using more latent groups improves the results but increases the computational effort significantly.

For all experiments, we used GPUs of type NVIDIA Tesla V100 with 32GB of memory and trained with an AdaMax optimizer [70]. We applied a cosine annealing schedule for the learning rate [71] starting at 0.01 and ending at 0.0001.

B Derivation of Bayesian Aggregation

We use two related encoders to learn a latent observation $\boldsymbol{\mu}_i = \text{enc}_\mu(\boldsymbol{x}_i, \boldsymbol{y})$ with its corresponding variance values $\boldsymbol{\sigma}_i = \text{enc}_\sigma(\boldsymbol{x}_i, \boldsymbol{y})$.

Assuming a factorized Gaussian prior distribution in the latent space $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{\boldsymbol{z},0}, \text{diag}(\boldsymbol{\sigma}_{\boldsymbol{z},0}))$, we can derive the factorized posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \text{diag}(\boldsymbol{\sigma}_z))$ in closed form using standard Gaussian conditioning [66] following [75]

$$\boldsymbol{\sigma}_{\boldsymbol{z}}^2 = \left[(\boldsymbol{\sigma}_{\boldsymbol{z},0}^2)^\ominus + (\boldsymbol{\sigma}_i^2)^\ominus\right]^\ominus, \tag{9}$$

$$\boldsymbol{\mu}_{\boldsymbol{z}} = \boldsymbol{\mu}_{\boldsymbol{z},0} + \boldsymbol{\sigma}_{\boldsymbol{z},0}^2 \odot (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{\boldsymbol{z},0}) \oslash \boldsymbol{\sigma}_i^2 \tag{10}$$

where $\ominus$ denotes element-wise inversion, $\odot$ denotes element-wise multiplication, and $\oslash$ denotes element-wise division.

C Derivation of FusionVAE's ELBO

We start with the following KL divergence between the approximate posterior and the real posterior,

$$\text{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{y})||p_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})) \geq 0. \tag{11}$$

Next, we apply the Bayes's theorem to obtain

$$-\int q_\theta(\boldsymbol{z}|\boldsymbol{y}) \log \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{y}|\boldsymbol{x}) q_\theta(\boldsymbol{z}|\boldsymbol{y})} dz \geq 0. \tag{12}$$

This leads to

$$\begin{aligned} -\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})] - \text{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{y})||p(\boldsymbol{z}|\boldsymbol{x})) \\ + \int q_\theta(\boldsymbol{z}|\boldsymbol{y}) \log p(\boldsymbol{y}|\boldsymbol{x}) dz \geq 0. \end{aligned} \tag{13}$$

The term $\log p(\boldsymbol{y}|\boldsymbol{x})$ can be moved out from the third integral component, and leaves the integral becoming 1. Finally, we obtain the ELBO of the conditional log-likelihood

$$\log p(\boldsymbol{y}|\boldsymbol{x}) \geq \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})] + \text{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{y})||p(\boldsymbol{z}|\boldsymbol{x})). \tag{14}$$

D Ablation Studies

This is a supplement for the aggregation ablation study in Sec. 6.3. In Tab. 5, we saw that the average NLL of all experiments using mean and max aggregation methods are similar. Fig. 6 shows the corresponding qualitative results. However, even though the NLL is very similar, the results of the aggregation of all features (MaxAggAll and MeanAggAll) are much more blurry than the results of the aggregation with addition (MaxAggAdd and MeanAggAdd). This is in conformity with the MSE_{min} results. It indicates that the NLL alone is not always the best metric to assess the visual closeness to real faces. When carefully examining the images of the addition aggregations, you could argue that the predictions with zero input images look slightly more realistic for max aggregation while for three input images, mean aggregation seems to be marginally better. This again confirms the validity of the MSE_{min} results even though the NLL results are also in accordance for this comparison.

Fig. 6: Prediction results of the different aggregation methods on FusionCelebA for zero to three input images.

E Statistic Significance of the Results

All experiments for this publications were carefully designed and optimized so that the training procedures are stable and lead to reproducible results. However, the data processing pipelines introduce randomness which lead to non-deterministic training outcomes due to multi-GPU training. We therefore ran every experiment three times and reported the results of the best training in Sec. 6. In Tabs. 7 to 12 we provide the means and variances of the three training runs.

	0	1	2	3	avg
CVAE	17.67 ± 0.10	15.11 ± 0.07	14.19 ± 0.08	13.71 ± 0.07	15.27 ± 0.02
CVAE+S	18.45 ± 0.02	14.64 ± 0.06	$\mathbf{13.22} \pm 0.03$	$\mathbf{12.32} \pm 0.02$	14.81 ± 0.03
FusionVAE	$\mathbf{15.91} \pm 0.03$	$\mathbf{14.13} \pm 0.07$	13.64 ± 0.09	13.41 ± 0.10	$\mathbf{14.34} \pm 0.07$

Table 7: Mean and standard deviation of the FusionMNIST NLL results in 10^{-2} BPD. The best results are printed in bold.

	0	1	2	3	avg
FCN	10.84 ± 0.37	5.96 ± 0.11	6.02 ± 0.18	6.13 ± 0.25	7.38 ± 0.11
FCN+S	6.21 ± 0.65	3.79 ± 0.04	2.64 ± 0.07	1.88 ± 0.08	3.73 ± 0.22
CVAE	3.87 ± 0.03	1.76 ± 0.03	1.09 ± 0.03	0.83 ± 0.02	1.97 ± 0.03
CVAE+S	3.53 ± 0.06	1.77 ± 0.01	1.23 ± 0.04	1.02 ± 0.04	1.96 ± 0.01
FusionVAE	$\mathbf{3.14} \pm 0.01$	$\mathbf{1.04} \pm 0.06$	$\mathbf{0.77} \pm 0.04$	$\mathbf{0.67} \pm 0.03$	$\mathbf{1.47} \pm 0.03$

Table 8: Mean and standard deviation of the FusionMNIST MSE_{min} results in 10^{-2}. The best results are printed in bold.

	0	1	2	3	avg
CVAE	456.9 ± 9.42	289.1 ± 6.53	278.9 ± 4.51	270.3 ± 3.66	324.0 ± 5.17
CVAE+S	487.4 ± 27.45	355.4 ± 60.39	280.7 ± 35.12	230.9 ± 22.59	338.8 ± 22.99
FusionVAE	$\mathbf{251.0} \pm 2.06$	$\mathbf{222.3} \pm 3.71$	$\mathbf{226.8} \pm 3.16$	$\mathbf{224.0} \pm 3.36$	$\mathbf{231.0} \pm 2.05$

Table 9: Mean and standard deviation of the FusionCelebA NLL results in 10^{-2} BPD. The best results are printed in bold.

	0	1	2	3	avg
FCN	13.07 ± 0.87	20.00 ± 3.77	17.87 ± 3.42	15.56 ± 3.08	16.62 ± 2.48
FCN+S	11.94 ± 0.52	18.58 ± 7.02	14.90 ± 6.59	11.19 ± 5.39	14.15 ± 4.65
CVAE	8.70 ± 0.42	6.25 ± 2.16	4.33 ± 1.77	3.02 ± 1.37	5.58 ± 1.21
CVAE+S	9.87 ± 1.10	9.60 ± 3.34	7.30 ± 3.07	5.57 ± 2.64	8.09 ± 2.19
FusionVAE	$\mathbf{5.82} \pm 0.52$	$\mathbf{1.10} \pm 0.16$	$\mathbf{0.93} \pm 0.06$	$\mathbf{0.84} \pm 0.03$	$\mathbf{2.18} \pm 0.17$

Table 10: Mean and standard deviation of the FusionCelebA MSE_{min} results in 10^{-2}. The best results are printed in bold.

	0	1	2	3	avg
CVAE	25.27 ± 0.04	24.00 ± 0.25	22.98 ± 0.24	23.34 ± 0.19	23.90 ± 0.17
CVAE+S	26.12 ± 0.23	25.29 ± 0.27	24.27 ± 0.25	24.15 ± 0.20	24.97 ± 0.16
FusionVAE	$\mathbf{24.32} \pm 0.10$	$\mathbf{23.09} \pm 0.02$	$\mathbf{22.25} \pm 0.02$	$\mathbf{22.90} \pm 0.02$	$\mathbf{23.15} \pm 0.04$

Table 11: Mean and standard deviation of the FusionT-LESS NLL results in 10^{-2} BPD. The best results are printed in bold.

	0	1	2	3	avg
FCN	5.88 ± 0.04	3.32 ± 0.10	2.50 ± 0.09	1.96 ± 0.10	3.43 ± 0.06
FCN+S	8.83 ± 1.05	1.95 ± 0.14	1.28 ± 0.15	0.86 ± 0.14	3.26 ± 0.37
CVAE	5.49 ± 0.11	1.73 ± 0.22	0.95 ± 0.18	0.44 ± 0.07	2.18 ± 0.13
CVAE+S	4.87 ± 0.06	2.98 ± 0.29	2.06 ± 0.19	1.27 ± 0.09	2.81 ± 0.12
FusionVAE	$\mathbf{4.15} \pm 0.03$	$\mathbf{0.62} \pm 0.03$	$\mathbf{0.33} \pm 0.03$	$\mathbf{0.20} \pm 0.02$	$\mathbf{1.34} \pm 0.02$

Table 12: Mean and standard deviation of the FusionT-LESS MSE_{min} results in 10^{-2}. The best results are printed in bold.

F Reconstruction

Figs. 7 to 9 visualize the reconstruction outputs for all our datasets and architectures. For these results, the target image is always given as input. The first three rows of each figure show the reconstruction, when additionally three noisy or occluded input images are fed into the network.

The images show that our FusionVAE reconstructs the target images almost perfectly for all three datasets. On FusionMNIST, only the FCN does not manage to reconstruct the target images but shows blurry versions of them. We also see the same behavior for FusionCelebA and FusionT-LESS which underlines the importance of skip connections for this type of network. On FusionCelebA, we see that CVAE+S suffers from numeric instabilities causing colorful artifacts in some images. Omitting the skip connections here avoids that issue. On FusionT-LESS, all baseline methods create more or less blurry versions of the target image when just the target image is given. When inputting the occluded images in addition to the target image, the reconstruction is much better which shows that these networks have over-fitted to the task of removing occluded objects so that they cannot deal well with non-occluded images. In contrast, FusionVAE has the ability to reconstruct non-occluded input images very well.

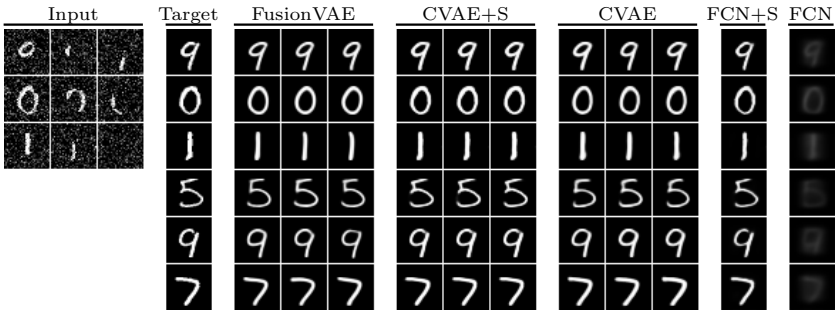

Fig. 7: Reconstruction results of the different architectures on FusionMNIST.

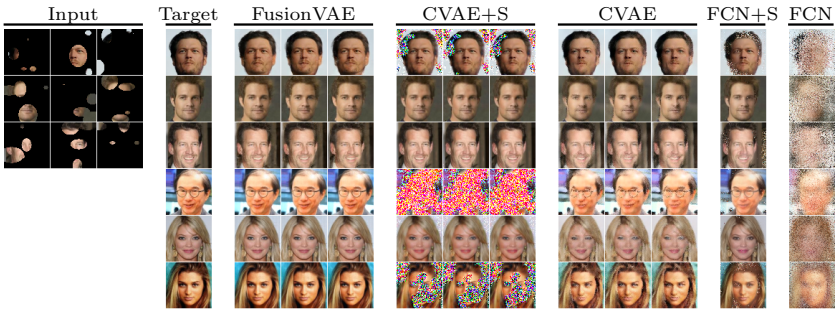

Fig. 8: Reconstruction results on FusionCelebA.

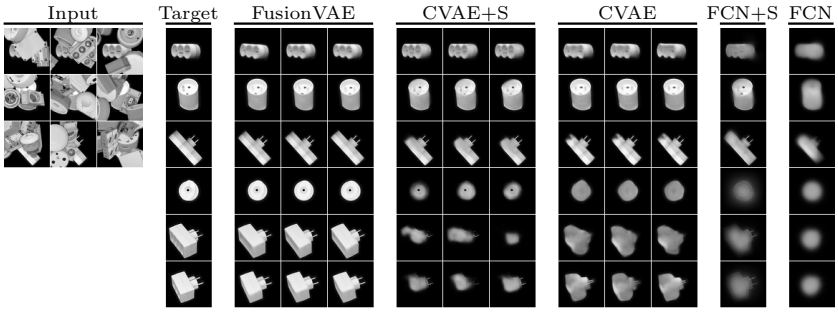

Fig. 9: Reconstruction results on FusionT-LESS.

References

66. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
67. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR. pp. 1251–1258 (2017)
68. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
69. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456. PMLR (2015)
70. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: ICLR (May 2015)
71. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: ICLR (2017)
72. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. In: ICLR (2018)
73. Salimans, T., Karpathy, A., Chen, X., Kingma, D.P.: PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. In: ICLR (2017)
74. Vahdat, A., Kautz, J.: NVAE: A deep hierarchical variational autoencoder. In: NeurIPS. vol. 33, pp. 19667–19679 (2020)
75. Volpp, M., Flürenbrock, F., Grossberger, L., Daniel, C., Neumann, G.: Bayesian context aggregation for neural processes. In: ICLR (2020)