Jointly Trained Variational Autoencoder for Multi-Modal Sensor Fusion

1st Timo Korthals
Bielefeld University
Cognitronics & Sensor Systems
Bielefeld, Germany
tkorthals@cit-ec.uni-bielefeld.de

2nd Marc Hesse
Bielefeld University
Cognitronics & Sensor Systems
Bielefeld, Germany
mhesse@cit-ec.uni-bielefeld.de

3rd Jürgen Leitner
Australian Centre for Robotic Vision
Queensland University of Technology
Brisbane, Australia
j.leitner@qut.edu.au

4th Andrew Melnik
Bielefeld University
Neuroinformatics Group
Bielefeld, Germany
anmelnik@techfak.uni-bielefeld.de

5th Ulrich Rückert
Bielefeld University
Cognitronics & Sensor Systems
Bielefeld, Germany
rueckert@cit-ec.uni-bielefeld.de

***Abstract*—This work presents the novel multi-modal Variational Autoencoder approach M^{2}VAE which is derived from the complete marginal joint log-likelihood. This allows the end-to-end training of Bayesian information fusion on raw data for all subsets of a sensor setup. Furthermore, we introduce the concept of in-place fusion – applicable to distributed sensing – where latent embeddings of observations need to be fused with new data. To facilitate in-place fusion even on raw data, we introduced the concept of a re-encoding loss that stabilizes the decoding and makes visualization of latent statistics possible. We also show that the M^{2}VAE finds a coherent latent embedding, such that a single naïve Bayes classifier performs equally well on all permutations of a bi-modal Mixture-of-Gaussians signal. Finally, we show that our approach outperforms current VAE approaches on a bi-modal MNIST & fashion-MNIST data set and works sufficiently well as a preprocessing on a tri-modal simulated camera & LiDAR data set from the Gazebo simulator.**

***Index Terms*—Multi-Modal Fusion, Deep Generative Model, Variational Autoencoder**

I. Introduction

Deep multi-modal generative models by means of *Variational Autoencoders* (VAE) are an upcoming research topic for sensor fusion and represent a subcategory of deep neuronal networks that facilitate a variational Bayes approach [1]–[4]. VAEs have a considerable impact on the field of data-driven learning of generative models as they tend to learn the inverse and forward models from observations in an unsupervised fashion. Furthermore, recent investigations have shown the fruitful applicability to zero-shot domain transfer in *deep reinforcement learning* (DRL) and bi-directional exchange of multi-modal data [3], [5].

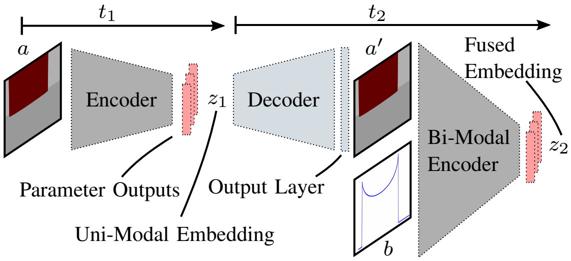

Fig. 1. Fusion of a camera and LiDAR perception from the Autonomous Mini-Robot (AMiRo) [11], sensing a red cylinder, to a single latent embedding z_2. The interaction is shown between the multi-modal encoder (right) facilitating sensor fusion of the formerly encoded latent embedding z_1 (left) and the new observation b. First (t_1), the color determining perception a by the camera is encoded to $z_1 \in \mathcal{Z}$. Second (t_2), a shape determining perception b – e.g. via a LiDAR – is fused via decoding z_1 to a' and encoding both into the bi-modal embedding $z_2 \in \mathcal{Z}$. The embeddings are determined by the parameter layers' outputs (e.g. mean and variance for VAEs trained with Gaussian sampling).

VAEs encode observations into latent space features that are (ideally) linearly separable [6]. More intriguingly is the ability to discover the joint posterior and likelihood models [3]. This facilitates sensor fusion by neuronal networks which obey variational Bayes methods. However, a framework that learns coherent posterior models between all subsets in a sensor setup remains unknown to the authors. Such a framework could handle sensory dropout during operation and therefore stabilize and lean subsequent classifying and reinforcement learning approaches, which commonly have to learn dropout during training (e.g. [7]). Furthermore, it may give an end-to-end solution to inverse sensor modeling relying on binary Bayes filters [8], [9] and may overcome their limitations regarding multi-modal fusion [8], [10].

This contribution exploits VAEs to circumvent the simplifying assumption of conditionally independent measurements for distributed estimation (c.f. [12]) by following a data-driven approach which models the full posterior distribution in a multi-

This research was supported by the Ministry of Economy, Innovation, Digitization and Energy of the State of North Rhine-Westphalia (MWIDE) within the Leading-Edge Cluster 'Intelligent Technical Systems OstWestfalenLippe (it's OWL)', supervised by Projektträger Jülich (PtJ), the Federal Ministry of Education and Research (57388272), and by 'CITEC' (EXC 277) at Bielefeld University which is funded by the German Research Foundation (DFG). The responsibility for the content of this publication lies with the author.

modal, in-place, posterior fusion approach based on VAEs. This approach is applicable in distributed active-sensing tasks (c.f. [13]). Compressed representations – e.g. the latent space's embedding z – of an object's observations $\mathcal{M}'$ are efficiently transmitted between all sensing agents and independently updated as follows: As depicted in Fig. 1, $z_1 \in \mathcal{Z}$ can be unfolded to the original observation using the VAE's decoder networks and combined with any new observation b to update the information in-place $z_1 \rightarrow z_2 \in \mathcal{Z}$.

We present a novel approach to build and train a multi-modal VAE (M^{2}VAE), which models the posterior (i.e. encoders or inverse model) and likelihood (i.e. decoder or forward model) of all combinations of modalities, that comprises the complete marginal joint log-likelihood without loss of generality. Furthermore, we propose a novel objective to maintain re-encoded embeddings (i.e. observation → encoding → decoding → encoding → ...) which is necessary to facilitate our proposed fusion approach.

Section II comprises the related work on multi-modal VAEs. Our approach is explained in Sec. III. To investigate the characteristic of the latent space $\mathcal{Z}$ as well as the quantitative features for existing multi-modal VAEs, we consider explainable data sets consisting out of an entangled MNIST & fashion-MNIST and simulated LiDAR & camera readings from the Gazebo simulator, which are described in Sec. IV and evaluated in Sec. V. Finally, we conclude our work in Sec. VI.

II. RELATED WORK

Variational Autoencoder (VAE) combine neural networks with variational inference to allow unsupervised learning of complicated distributions according to the graphical model shown in Fig. 2 a). A D_{a}-dimensional observation a is modeled in terms of a D_{z}-dimensional latent vector z using a probabilistic decoder $p_{\theta_{\mathrm{a}}}(z)$ with parameters θ. To generate the corresponding embedding z from observation a, a probabilistic encoder network with $q_{\phi_{\mathrm{a}}}(z)$ is being provided which parametrizes the posterior distribution from which z is sampled. The encoder and decoder, given by neural networks, are trained jointly to bring a close to an a' under the constraint that an approximate distribution needs to be close to a prior $p(z)$ and hence inference is basically learned during training.

The specific objective of VAEs is the maximization of the marginal distribution $p(a) = \int p_\theta(a|z)p(z)\,\mathrm{d}z$. Because this distribution is intractable, the model is instead trained via *stochastic gradient variational Bayes* (SGVB) by maximizing the *evidence lower bound* (ELBO) $\mathcal{L}$ of the marginal log-likelihood $\log p(a) := L_{\mathrm{a}}$ as

$$L_{\mathrm{a}} \geq \mathcal{L}_{\mathrm{a}} = \underbrace{-\,\mathrm{D}_{\mathrm{KL}}(q_\phi(z|a)\|p(z))}_{\text{Regularization}} + \underbrace{\mathbb{E}_{q_\phi(z|a)}\log(p_\theta(a|z))}_{\text{Reconstruction}} . \tag{1}$$

This approach proposed by [14] is used in settings where only a single modality a is present in order to find a latent encoding z (c.f. Fig. 2 a)).

In the following chapters, we briefly comprise related work by means of multi-modal VAEs. Further, we stress the concept of two joint multi-modal approaches to derive the later proposed M^{2}VAE.

A. Multi-Modal Auto Encoder

Given a set of modalities $\mathcal{M} = \{\text{a,b,c},\ldots\}$, multi-modal variants of *Variational Auto Encoder*s (VAE) have been applied to train generative models for multi-directional reconstruction (i.e. generation of missing data) or feature extraction. Variants are *Conditional VAEs* (CVAE) and *Conditional Multi-Modal AEs* (CMMA), with the lack in bi-directional reconstruction (c.f. [15], [16]). BiVCCA by [1] trains two VAEs together with interacting inference networks to facilitate two-way reconstruction with the lack of directly modeling the joint distribution. Models, that are derived from the *Variation of Information* (VI) with the objective to estimate the joint distribution with the capabilities of multi-directional reconstruction were recently introduced by [3]. [4] introduce another objective for the bi-modal VAE, which they call the triplet ELBO (tVAE). Furthermore, multi-modal stacked AEs are a variant of combining the latent spaces of various AEs (c.f. [17], [18]) which can also be applied to the reconstruction of missing modalities ([2], [19]). However, while [3] and [4] argue that training of the full multi-modal VAE is intractable, because of the $2^{|\mathcal{M}|}{-}1$ modality subsets of inference networks, we show that training the full joint model estimates the most expressive latent embeddings.

1) Joint Multi-Modal Variational Auto Encoder: When more than one modality is available, e.g. a and b as shown in Fig. 2 a), the derivation of the ELBO $\mathcal{L}_{\mathrm{J}}$ for a marginal joint log-likelihood $\log p(a) := L_{\mathrm{J}}$ is straight forward:

$$L_{\mathrm{J}} \geq \mathcal{L}_{\mathrm{J}} = \underbrace{-\,\mathrm{D}_{\mathrm{KL}}(q_{\phi_{\mathrm{ab}}}(z|a{,}b)\|p(z))}_{\text{Regularization}} + \tag{2}$$

$$\underbrace{\mathbb{E}_{q_{\phi_{\mathrm{ab}}}(z|a,b)}\log(p_{\theta_{\mathrm{a}}}(a|z))}_{\text{Reconstruction wrt. } a} + \underbrace{\mathbb{E}_{q_{\phi_{\mathrm{ab}}}(z|a,b)}\log(p_{\theta_{\mathrm{b}}}(b|z))}_{\text{Reconstruction wrt. } b} . \tag{3}$$

However, given Eq. 3 it is not clear how to perform inference if the dataset consists of samples lacking from modalities (e.g. for samples i and k: $(a_i,\varnothing)$ and $(\varnothing,b_k)$). [2] propose training of a bimodal deep auto encoder using an augmented dataset with additional examples that have only a single-modality as input. We, therefore, name the resulting model of Eq. 3 *joint multi-modal VAE-Zero* (JMVAE-Zero).

2) Joint Multi-Modal Variational Auto Encoder from Variation of Information: While the former approach cannot directly be applied to missing modalities, [3] propose a *joint multi-modal VAE* (JMVAE) that is trained via two uni-modal encoders and a bi-modal en-/decoder which share one objective function derived from the *Variation of Information* (VI) of the marginal conditional log-likelihoods $\log p(a|b)p(b|a) =: L_{\mathrm{M}}$ by optimizing the ELBO $\mathcal{L}_{\mathrm{M}}$:

$$L_{\mathrm{M}} \geq \mathcal{L}_{\mathrm{M}} \geq \mathcal{L}_{\mathrm{J}} - \tag{4}$$

$$\underbrace{\mathrm{D}_{\mathrm{KL}}(q_{\phi_{\mathrm{ab}}}(z|a{,}b)\|q_{\phi_{\mathrm{b}}}(z|b))}_{\text{Unimodal PDF fitting of encoder b}} - \underbrace{\mathrm{D}_{\mathrm{KL}}(q_{\phi_{\mathrm{ab}}}(z|a{,}b)\|q_{\phi_{\mathrm{a}}}(z|a))}_{\text{Unimodal PDF fitting of encoder a}} \tag{5}$$

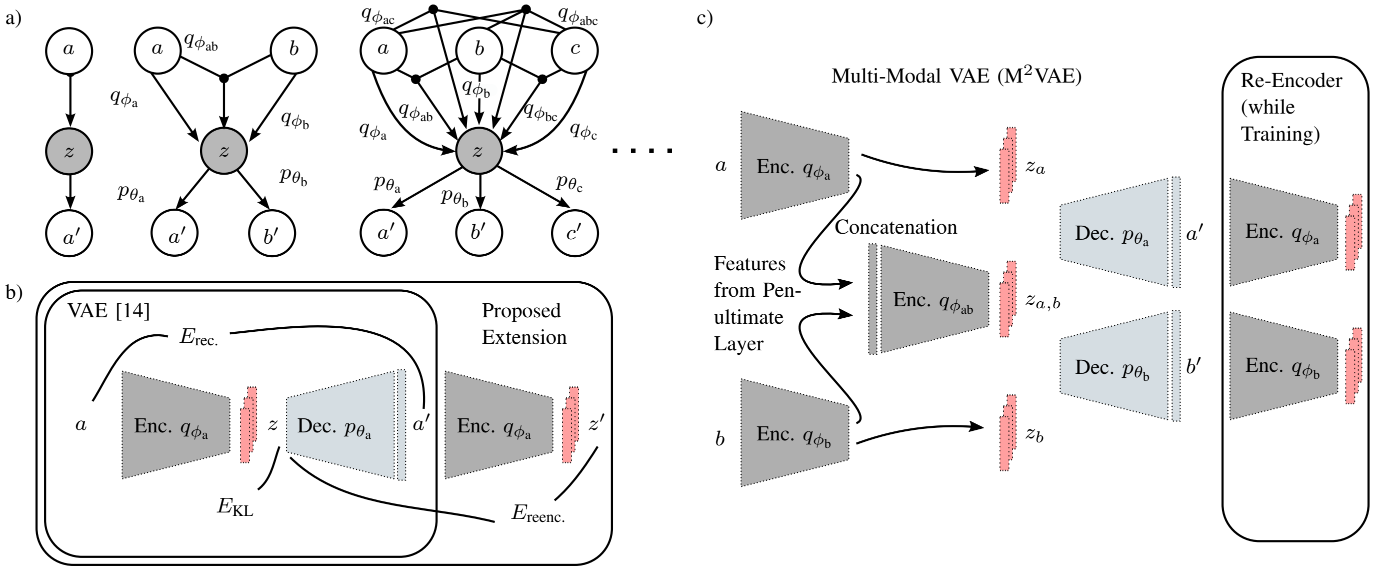

Fig. 2. a) Evolution of full uni-, bi-, and tri-modal VAEs comprising all modality permutations. b) Extension of the standard VAE to facilitate immutable re-encoding of the latent embedding z. c) Multi-modal VAE (MVAE) realization in the bi-modal case with proposed re-encoding during training from b). To maintain stability during training and to keep the number of parameters tractable, outputs from the penultimate layers (i.e. before the linear, distribution parameterizing layers) are taken as input values for multi-modal encoders.

Therefore, uni-modal encoders are trained, so that their distributions q_{ϕ_a} and q_{ϕ_b} are close to a multi-modal encoder $q_{\phi_{\text{ab}}}$ in order to build a coherent posterior distribution. The introduced regularization by [3] puts learning pressure on the uni-modal encoders just by the distributions' shape, disregarding reconstruction capabilities and the prior $p(z)$. Furthermore, one can show that deriving the ELBO from the VI for a set of $\mathcal{M}$ observable modalities, always leads to an expression of the ELBO that allows only training of $\widetilde{\mathcal{M}} = \{m | m \in \mathcal{P}(\mathcal{M}), |m| = |\mathcal{M}| - 1\}$ modality combinations. This leads to the fact that for instance in a tri-modal setup, as shown in Fig. 2 a), one can derive and train three bi-modal encoders, but no uni-modal ones.

III. VAE Fusion Approach

While the objective of [1], [2], [3], and [4] is to exchange modalities bi-directionally (e.g. $a \rightarrow b'$), our primary concern is twofold: First, find an expression to jointly train all $2^{|\mathcal{M}|}-1$ permutations of modality encoders (c.f. Sec. III-A). Second, add an additional objective that ensures immutability while re-encoding observations to facilitate in-place sensor fusion (c.f. Sec. III-B).

A. Multi-Modal Variational Autoencoder

By successively applying logarithm and Bayes rules, we derive the ELBO for the multi-modal VAE (M^{2}VAE) as follows: First, given the independent set of observable modalities $\mathcal{M} = \{\text{a,b,c}, \ldots\}$, its marginal log-likelihood $\log p(\mathcal{M}) =: L_{\text{M}^2}$ is multiplied by the cardinality of the set as the neutral element $1 = {}^{|\mathcal{M}|}/_{|\mathcal{M}|}$. Second, applying logarithm multiplication rule, the nominator is written as the argument's exponent. Third, Bayes rule is applied to each term wrt. the remaining observable modalities to derive their conditionals. Further, we bootstrap the derivation technique in a bi-modal (c.f. [20] for tri-modal) case to illustrate the advantages. Excessively applying the scheme until convergence of the mathematical expression leads for the bi-modal set $\mathcal{M} = \{\text{a,b}\}$ to the term in Eq. 7.

$$L_{\text{M}^2} = \tfrac{2}{2} \log p(a,b) = \tfrac{1}{2} \log p(a,b)^2 \tag{6}$$

$$= \tfrac{1}{2} \log p(a,b) p(a,b) \tfrac{1}{2} \log p(b) p(a|b) p(b|a) p(a) \tag{7}$$

$$= \tfrac{1}{2} (\log p(a) + \log p(b|a) + \log p(a|b) + \log p(b)) \tag{8}$$

$$\overset{\text{Eq. 1, 5}}{=} \tfrac{1}{2} (L_{\text{a}} + L_{\text{M}} + L_{\text{b}}) \tag{9}$$

This term can be written as inequality wrt. each ELBO of the marginals L_{a}, L_{b} and conditionals L_{M}:

$$2L_{\text{M}^2} \geq 2\mathcal{L}_{\text{M}^2} = \mathcal{L}_{\text{a}} + \mathcal{L}_{\text{b}} + \mathcal{L}_{\text{M}} = \tag{10}$$

$$- \mathrm{D}_{\text{KL}}(q_{\phi_{\text{a}}}(z|a) \| p(z)) + \mathbb{E}_{q_{\phi_{\text{a}}}(z|a)} \log(p_{\theta_{\text{a}}}(a|z)) \tag{11}$$

$$- \mathrm{D}_{\text{KL}}(q_{\phi_{\text{b}}}(z|b) \| p(z)) + \mathbb{E}_{q_{\phi_{b}}(z|b)} \log(p_{\theta_{\text{b}}}(b|z)) \tag{12}$$

$$+ \mathbb{E}_{q_{\phi_{\text{ab}}}(z|a,b)} \log(p_{\theta_{\text{a}}}(a|z)) + \mathbb{E}_{q_{\phi_{\text{ab}}}(z|a,b)} \log(p_{\theta_{\text{b}}}(b|z)) \tag{13}$$

$$- \mathrm{D}_{\text{KL}}(q_{\phi_{\text{ab}}}(z|a,b) \| p(z)) \tag{14}$$

$$- \mathrm{D}_{\text{KL}}(q_{\phi_{\text{ab}}}(z|a,b) \| q_{\phi_{\text{a}}}(z|a)) - \mathrm{D}_{\text{KL}}(q_{\phi_{\text{ab}}}(z|a,b) \| q_{\phi_{\text{b}}}(z|b)). \tag{15}$$

Equation 10 is substituted by all formerly derived ELBO expressions lead to the combination of the uni-modal VAEs wrt. a and b (c.f. Eq. 11 and 12) and the JMVAE comprising the VAE wrt. the joint modality ab (c.f. Eq. 13 and 14) and mutual latent space (c.f. Eq. 15). Equation 11 and 12 have the effect that their regularizers care about the uni-modal distribution to deviate not too much from the common prior while their reconstruction term shapes the underlying embedding of the mutual latent space. A network configuration, comprising the three encoder and two decoder networks from Eq. 10, is depicted in Fig. 2 c). It is worth mentioning that one can apply the concept of β-VAE (c.f. [5], [21], [22]) to the regularizers via single scalar β_m. However, while β-VAE have the property to disentangle the latent space, our main concern is the balance between the input and latent space using a constant normalized factor $\beta_{\text{norm}} = \beta_m {}^{D_m}/_{D_{\text{z}}} \forall m \in \mathcal{P}\left(\mathcal{M}\right)$.

If the derivation is applied to the log-likelihood $L_{\text{M}^2{}_{\mathcal{M}}}$ of a set $\mathcal{M}$, one can show that it results into a recursive form consisting of JMVAEs' and M^{2}VAEs' log-likelihood terms

$$L_{\text{M}^2{}_{\mathcal{M}}} = \frac{1}{|\mathcal{M}|}\left(L_{\text{M}_{\mathcal{M}}} + \sum_{\widetilde{m}\in\widetilde{\mathcal{M}}} L_{\text{M}^2{}_{\widetilde{m}}}\right) \tag{16}$$

$$\geq \frac{1}{|\mathcal{M}|}\left(\mathcal{L}_{\text{M}_{\mathcal{M}}} + \sum_{\widetilde{m}\in\widetilde{\mathcal{M}}} \mathcal{L}_{\text{M}^2{}_{\widetilde{m}}}\right) =: \mathcal{L}_{\text{M}^2{}_{\mathcal{M}}}. \tag{17}$$

While the derivation of Eq. 17 is given in [20], the properties are as follows:

- the M^{2}VAE consist out of $2^{|\mathcal{M}|} - 1$ encoders and $|\mathcal{M}|$ decoders comprising all modality combinations
- while it also allows the bi-directional exchange of modalities, it further allows the setup of arbitrary modality combinations having 1 to $|\mathcal{M}|$ modalities
- subsets of minor cardinality are weighted less and have a therefore minor impact in shaping the overall posterior distribution (vice versa, the major subsets dominate the shaping, and the minor sets adapt to it)
- all encoder/decoder networks can jointly be trained using SGVB

B. In-Place Sensor Fusion

This section introduces the concept of in-place sensor fusion, that updates an existing embedding z to z^*, using multi-modal VAEs as follows:

$$q_{\phi_{m\cup\mathcal{M}'}}(z^*|m,\text{f}(\mathcal{M}')) \quad \text{with} \quad \text{f}(\mathcal{M}') = \bigcup_{m' in \mathcal{M}'} p_{\theta_{m'}}(m'|z). \tag{18}$$

However, a necessary requirement of Eq. 18 is, that auto re-encoding (i.e. $z \rightarrow z$ via $q_{\phi_{\mathcal{M}'}}(z|\mathcal{M}')$) does not manipulate the information represented by z in an unrecoverable way (e.g. label-switching). One may assume that VAEs tend to have a natural denoising characteristic (despite the explicit denoising Autoencoders) which should re-encode any z in a better version of its own by means of the reconstruction loss wrt. the observation. Surprisingly, this behavior only holds for linear separable observations as discussed later in Sec. V. For non-separable data, the common VAE tend to re-encode any observation to the priors mean and thus, changes the initial information fundamentally. Similar observations were already made by [23] which contradict the basic assumption of in-place sensor fusion.

To maintain stability and immutability of the encoding during re-encoding, we propose a new training objective by adding a re-encoding loss $E_{\text{reenc.}}$ to the common VAE objective (c.f. Fig. 2 b)) This results in the new loss term comprising the reconstruction losses $E_{\text{rec.}}$, prior and mutual loss E_{KL}, and the proposed re-encoding loss $E_{\text{reenc.}}$:

$$E = E_{\text{rec.}} + E_{\text{KL}} + \alpha E_{\text{reenc.}}. \tag{19}$$

$E_{\text{reenc.}}$ can be any loss function or metric that compares either the sampled encoding or distribution parameters. The parameter α scales re-encoding loss to leverage its influence in contrast to the reconstruction and prior losses.

IV. Data Sets

[24] state that Hebbian learning relies on the fact that the same objects are continuously transformed to their nearest neighbor in the observable space. [21] adopted this approach to their assumptions, that this notion can be generalized within the latent manifold learning. Further, neither a coherent manifold nor a proper factorization of the latent space can be trained if these assumptions are not fulfilled by the dataset. In summary, this means that observed data has to have the property of continuous transformation wrt. to their properties (e.g. position and shape of an object), such that a small deviation of the observations results in proportional deviations in the latent space. We adopt this assumption for multi-modal data sets where observations should correlate if the same quantity is observed, such that a small deviation in the common latent representation between all modalities conducts a proportional impact in all observations. This becomes an actual fundamental requirement for any multi-modal data set, as correlation and coherence are within the objective of multi-modal sensor fusion. However, quantities may be partially observable, so that the complete state of an observation can be obtained via complementary fusion.

It is quite common in the multi-modal VAE community to model a bi-modal data set as follows (c.f. [1]–[4]): The first modality a denotes the raw data and b denotes the label (e.g. the digits' images and labels as one-hot vector wrt. the MNIST data set). This is a rather artificial assumption and only sufficient when the objective is within a semi-supervised training framework. Real multi-modal data does not show this behavior as there are commonly multiple raw data inputs. While only complex multi-modal data sets of heterogeneous sensor setups exist (c.f. [25]), which makes an explainable evaluation for our approach futile, we generate various data sets for evaluation on our own: First, we evaluate a bi-modal, 10 class *Mixture of Gaussians* (MoG) data set (c.f. Fig. 3) that is ideally separable by a Gaussian naïve Bayes classifier. Modality a realizes the class observations on a two-dimensional grid with each class noise being $\sigma_{\text{a}} = .06$, while modality b observes a projection on a unit cycle with $\sigma_{\text{b}} = .1$. Furthermore, to investigate the complementary fusion capabilities, the mean values of a's observations are consolidated for the classes $(5{,}6{,}7)$ and $(0{,}8)$, while b observes the consolidated classes $(0{,}9)$. However, every class is ideally separable if both modalities are observed.

While naïve consolidation of non-correlated data sets does not meet the conditions of data continuity and correlation, as discussed earlier, we secondly consolidate the MNIST and

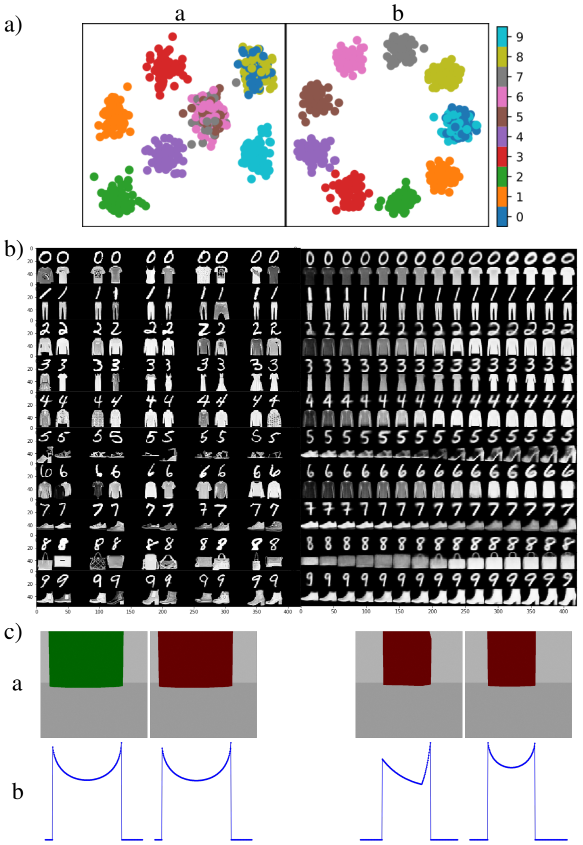

Fig. 3. a) MoG input signals with for the modalities a and b. The depicted observations are sampled for the corresponding modality for each class. b) Comparison of standard non-correlated MNIST and f-MNIST data set (left) and our proposed MNIST-E data set (right). c) Observations of objects via camera and LiDAR in the Gazebo simulation with ambiguous observations wrt. shape (left) and color (right). Reflectance measurements (i.e. modality c) is missing due to low dimensionality.

fashion-MNIST data set by sampling from superimposed latent spaces of various uni-modal trained conditional VAEs. This approach allows the generation of the bi-modal data sets, i.e. MNIST-E, from the distinct and disconnected uni-modal data sets.

Furthermore, we investigate a tri-modal data set collected via the *Autonomous Mini-Robot* [26] simulator comprising a camera, LiDAR, and reflectance sensor. Assuming closed world conditions, only primitive objects with the attributes $color \in \{\text{red}, \text{green}\}$, $shape \in \{\text{cylindric}, \text{cubic}\}$, and $reflectance \in \{\text{mat}, \text{shiny}\}$ exist, which results in $2^3 = 8$ objects. Therefore, every object is only assignable to one class in a classification task, if and only if every attribute is sensed.

V. Experiments

We apply the datasets explained in Sec. IV to test and depict the capabilities of the M^{2}VAE. First, we evaluate complementary fusion property on the MoG data set in Sec. V-A. Second, we investigate the proposed extended VAE objective with the additional re-encoding loss term briefly on the MNIST and MoG data set in Sec. V-B, which also introduces a novel visualization technique for latent spaces. Third, the more complex data sets MNIST-E (without in-place fusion) and LiDAR/camera (with in-place fusion) are evaluated in Sec. V-C and V-D.

Various VAEs are compared qualitatively, by visualizing the latent space, and quantitatively by performing lower bound tests $\mathcal{L}_{\widetilde{\mathcal{M}}}$ for every subset $\widetilde{\mathcal{M}} \subseteq \mathcal{M}$ wrt. to the decoding of all modalities $p_{\theta_{\mathcal{M}}}$:

$$\mathcal{L}_{\widetilde{\mathcal{M}}} = \mathbb{E}_{q_{\phi_{\widetilde{\mathcal{M}}}}(z|\widetilde{\mathcal{M}})} \log \frac{p_{\theta_{\mathcal{M}}}(\mathcal{M}|z)p(z)}{q_{\phi_{\widetilde{\mathcal{M}}}}\left(z|\widetilde{\mathcal{M}}\right)} \tag{20}$$

with $p(z) = \mathcal{N}(z; \mathbf{0}, \mathbf{I})$. However, we also qualitatively evaluate the latent space representation with the premise in mind, that a good generative model should not just generate good data, i.e. low reconstruction error, but also gives a good latent distribution of samples $z \in \mathcal{Z}$. All experiments are explained on the basis of the best performing network architectures evaluated via hyper parameter grid search. The corresponding code implementations and library for building M^{2}VAEs are publicly available.

Furthermore, we briefly argue about selection of a good β_{norm} parameter, since it controls the mutual connection between all encoders as well as the prior's impact We found that a high value (i.e. $\beta_{\text{norm}} \geq 1.$) put too much learning pressure on matching the mutual and – more importantly – the prior distributions which result in uninformative embeddings. To small values, on the other hand, leads to better reconstruction but also to non-coherent embeddings since the VAE is able to become an AE for $\beta_{\text{norm}} \rightarrow 0$. These findings makes the M^{2}VAE approach congruent to the behaviors of standard β-VAEs [6], so that a $\beta_{\text{norm}} \lessapprox 10^{-2}$ was chosen for training. It is worth mentioning that learning pressure via β_{norm} should be applied equally to all encoders so that they experience a similar learning impact wrt. the latent space.

Furthermore, we want to highlight that we trained the M^{2}VAE in advance to all classification.

A. MoG Experiment

The M^{2}VAE enforces its encoder networks q_{ϕ_*} inherently to approximate the same posterior distribution which can be seen by the strong coherence between all embeddings. In our depicted case, coherence means that the same observations lead to the same latent embedding: $q_{\phi_{\text{ab}}}(a,b) \approx q_{\phi_{\text{a}}}(a) \approx q_{\phi_{\text{b}}}(b)$. However, this property only holds for non-ambiguous observations. Observations made from classes which are not separable collapse to a common mean in the latent space, which is denoted for the uni-modal cases by (+) and (-). Furthermore, the embeddings also show an interesting behavior for samples from class (0): As this class is only ambiguously detectable in the uni-modal case, the encoder networks learn a separable, and therefore unambiguous, embedding if both modalities are present (denoted by (-)).

The depicted behaviors are also rendered by the ELBO $(-\mathcal{L})$, which was used as the objective for training the M^{2}VAE. This is an intriguing observation because while the samples are no longer separable (not even non-linearly) in latent space, the

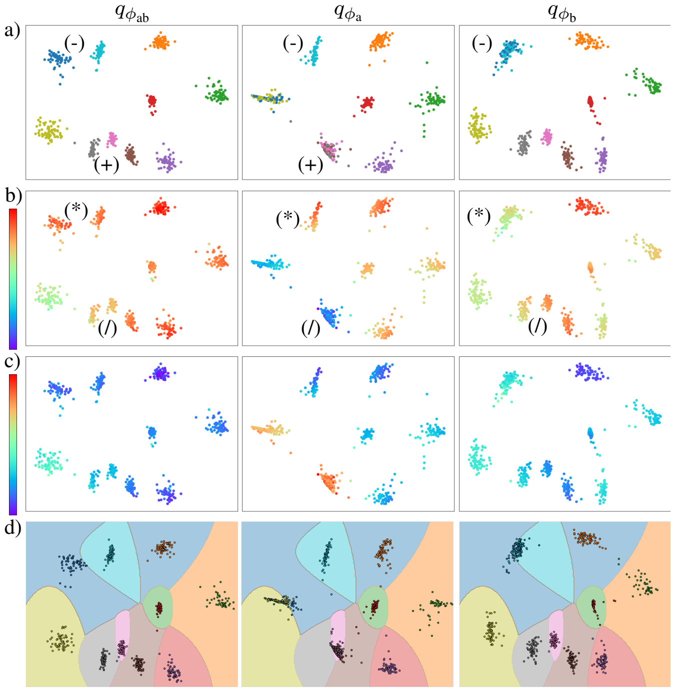

Fig. 4. 2-dimensional latent space embeddings of the bi-modal MoG test set. Plots from left to right show the embeddings of the bi-modal (a,b), uni-modal (a) and uni-modal (b) observation. Colorization: a) class labels (c.f. Fig 3), b) ELBO, c) D_{KL}, d) decision boundaries of a single naïve Bayes classifier.

ELBO for the observation goes down (c.f. $(*)$ and $(/)$) and gives, therefore, evidence about the embedding quality and information content. This insight might connect VAEs with the free-energy principle introduced by Friston [27] and might be fruitful in terms of epistemic (ambiguity resolving) tasks, where for instance an unsupervised learning approach could use the ELBO as a signal to learn epistemic action selection. However, while the ELBO is not accessible during inference, we also plotted the accessible Kullback–Leibler divergence (i.e. the prior loss $\mathrm{D}_{\mathrm{KL}}(q_{\phi_*}\|p(z))$), that is a value for the learned complexity of an observation [27]. This quantity, trained by our approach, behaves inversely to the ELBO as postulated by Fristen and will be investigated in prospective studies.

Figure 4 d) shows the interaction between the latent embeddings and a single naïve Bayes classifier, that was trained on these embeddings. As one logically needs three classifiers for classifying all permutations of observations ((a,b), (a), (b)), the MVAE projects all permutations such, that only one naïve classifier is necessary. This is an interesting insight because of the fact that this single classifier reaches the same classification rate (c.f. Table I) as three exclusive classifiers trained on the raw data. Furthermore, the ambiguous observations lie mainly on the decision boundaries of the classifier. Again, we want to highlight the M^{2}VAE, which performs the embedding, was trained in an unsupervised fashion. Therefore, we want to attribute this behavior to the feature that VAEs naturally project observations onto the prior by maintaining the sampling distribution. Both are – in our experiment – Gaussian and therefore seem to interact perfectly with a Gaussian naïve Bayes classifier, However, other multi-modal VAE approaches tend to learn non-coherent latent spaces which we attribute to the bad classification rates.

Enabling the possibility of using just a single classifier on a multi-modal sensor setup, that is susceptible to sensory dropout, is an outstanding feature of our approach. This could stabilize and lean future classifying and reinforcement learning approaches, which commonly learn dropout during training, such that they learn from the common and coherent latent space $\mathcal{Z}$.

TABLE I
CLASSIFICATION RATE, I.E. THE RATIO OF CORRECTLY CLASSIFIED SAMPLES TO THE TOTAL NUMBER OF SAMPLES, FOR THE NAÏVE BAYES CLASSIFICATION ON THE RAW AND ENCODED DATA.

input	Raw	Embedding		
		M^{2}VAE	JMVAE-Zero	tVAE
a&b	.99	**.99**	.99	.99
a	.71	**.71**	.63	.64
b	.90	**.90**	.09	.29

The M^{2}VAE for Fig. 4 were configured as follows: $q_{\phi_{\text{a}}}$ and $q_{\phi_{\text{b}}}$ have 2/input → 128/ReLU → 64/ReLU → two times 2/linear for mean and log-variance for Gaussian sampling → 64/ReLU → 128/ReLU → 2/sigmoid, batch size: 128, epochs: 400, $\beta_{\text{norm}} = .01$. $q_{\phi_{\text{ab}}}$ consists of a single 64/ReLU layer. The other VAEs are configured accordingly, but without the latent encoder $q_{\phi_{\text{ab}}}$. It is worth noticing that the VAEs do not learn the identity function, regardless of their high encoder fan-out ($D_{\text{a}} = 2$ vs. $D = 128$ of the first hidden layer), which we attribute to the sampling layers and the prior loss E_{KL} in the VAEs' bootlenecks.

B. Re-Encoding Experiment

The benefits of training a VAE via the proposed re-encoding loss approach are twofold: First, the re-encodings become nearly immutable and label switching can be suppressed (c.f. Fig. 5 a) and b)). The immutability of single $z_{\text{init.}} \in \mathcal{Z}$ are visualized as colorized perturbation by calculating the Euclidean distance z_{diff} between the encoding before and after (i.e. $z_{\text{reenc.}}$) re-encoding (c.f. Fig. 5 c)):

$$z_{\text{diff}} = \|z_{\text{init.}} - z_{\text{reenc.}}\|_2 \quad \text{with} \quad z_{\text{reenc.}} \sim q_\phi(z|p_\theta(z_{\text{init.}})). \quad (21)$$

While the common approach (left) without re-encoding loss shows high perturbation allover the embedding area (c.f. Fig. 5 a) vs. c)), the proposed approach (right) is nearly free off perturbation. Furthermore, the perturbation becomes higher outside of the embedding's area, indicated as red artifacts at the image borders, since the VAE was not trained to preserve the re-encodings in these areas. However, we take this as an indicator that the VAE does not cheat on the re-encodings by learning the identity function, which one might assume because of the high fan-out between z and z' (c.f. 2 b)) We attribute this feature to the reconstruction loss $E_{\text{rec.}}$ between z and z' that acts as a regularize between the re-encoding. Second, the latent spaces' statistics of the encoder networks can be visualized by traversing the latent space $\mathcal{Z}$ while obtaining the

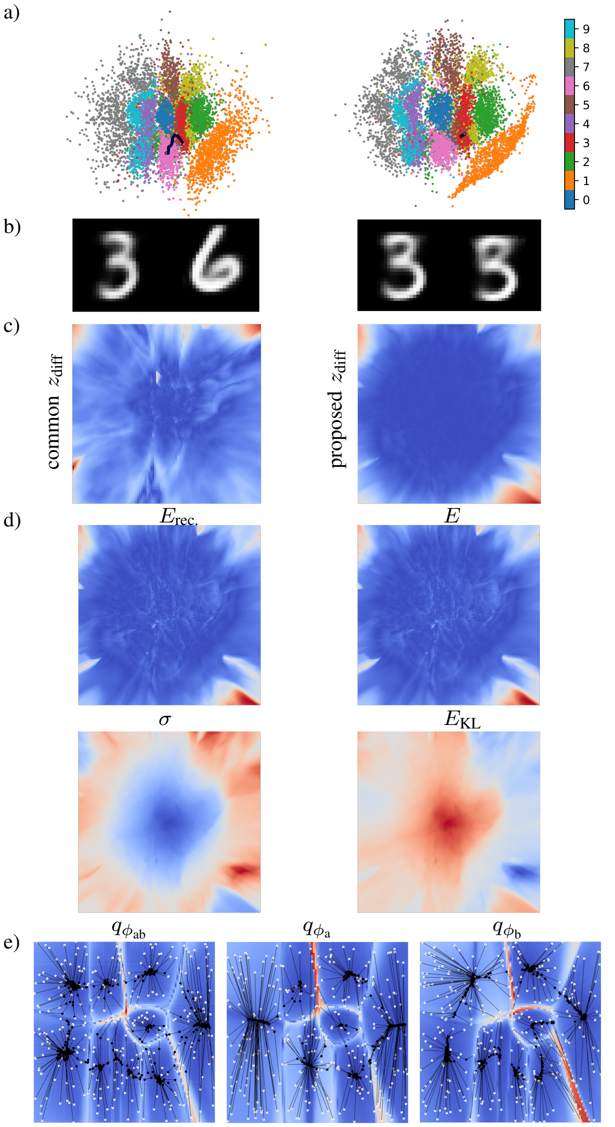

Fig. 5. a) – d) 2-dimensional latent space visualization of the MNIST test data using a VAE as proposed in [14] with Gaussian sampling layer and prior. a) Embeddings of a VAE's encoder q_ϕ trained via common (left) and the proposed loss (right) with $E_{\text{reenc.}} = \mathrm{D}_{\mathrm{KL}}\big(q_\phi(a)\|q_\phi(a')\big)$. Black trajectories indicate the initial encoding of a '3' and the terminal encoding after 400 re-encodings. b) Corresponding initial and final decodings. c) Qualitative difference z_{diff} between the initial and re-encoded z. d) Qualitative latent space statistics produced with auto re-encoding ($\sigma^2 = \sum_i \sigma_i^2$). c) and d) are heatmaps visualizations (blue i.e. low and red i.e. high values). d) Trajectory visualization of re-encoding using the jointly trained bi- and unimodal encoders, without proposed loss, on the MoG data with $E_{\text{rec.}}$ underlay.

output parameters of the encoder via re-encoding (c.f. Fig. 5 d)). Visualization of the latent space statistics gives intriguing insights to the behaviors of VAEs. As shown in Fig. 5 d), the encoder network tends to tie up the variances σ and therefore deviate from the prior, indicated by E_{KL}, where the encoder embeds observations into $\mathcal{Z}$. Furthermore, the reconstruction loss $E_{\text{rec.}}$ becomes higher at the vicinity of cluster boarders, where the encoder embeds poor or ambiguous observations. We also plotted the combined loss E (i.e. negative ELBO $\mathcal{L}$) for the sake of completeness which shows the same behavior, because of dominant reconstruction loss that is attributable to high input dimensionality ($D_{\text{a}} = 784$ vs. $D_{\text{z}} = 2$). However, this behavior is hardly recognizable by the slightly brighter filaments because of the already complex data set. It becomes much clearer for the MoG experiment which comprises linear separable data.

As mentioned earlier, for the linear separable MoG data set, the M^{2}VAE without the proposed loss does, in fact, tend to have a denoising characteristic which re-encodes any z in a refined version of its own by means of the reconstruction loss. This behavior is shown in Fig. 5 e) where we underlay the re-encoding trajectories with the reconstruction loss $E_{\text{rec.}}$. One can see naturally learned discrimination boarders of the latent space indicated by high losses which separate clusters' vicinities. Furthermore, initial z values are auto re-encoded which draw the trajectories along their path in latent space. The properties of the various VAE encoders q_{ϕ_*} during re-encoding show that every observation converges to a fixed-point, i.e. the corresponding clusters' mean values while performing descending steps on the loss manifold.

The VAEs for Fig. 5 a) – d) were both configured as follows: 784/input → 256/ReLU → 128/ReLU → two times 2/linear for mean and log-variance for Gaussian sampling → 128/ReLU → 256/ReLU → 784/sigmoid, batch size: 1024, epochs: 400, $\alpha = .01$ using warmup by [28] after 200 epochs.

C. MNIST-E Experiment

For this experiment, we estimated the ELBO by Eq. 20 to evaluate the performance of models JMVAE-Zero, tVAE, and M^{2}VAE. We chose the model wrt. to the evaluation in Fig. 4 with $\beta_{\text{norm}} = 10^{-2}$ for the given MNIST image resolution of $D_{\text{a}} = D_{\text{b}} = ||(28{,}28{,}1)||$ and $D_z = 10$. Since the MNIST-E set shares the same latent distribution for all subsets of modalities a (i.e. MNIST) and b (i.e. fashion-MNIST), we expect the VAE to learn equal evidence lower bounds for the uni- as well as the bi-modal observations. Table II shows quantitatively that the proposed M^{2}VAE reaches the highest ELBO value, as well as it meets almost meet the expectations of learning equal ELBOs. The other VAEs deviate from that expectation which we attribute the simplifications in their training objectives.

TABLE II
ELBO TEST FOR UNI- AND MULTI-MODAL VAEs (HIGHER IS BETTER).

	$\mathcal{L}_{\text{a,b}}$	$\mathcal{L}_{\text{a}}$	$\mathcal{L}_{\text{b}}$
M^{2}VAE	$\mathbf{-10.75}$	$\mathbf{-10.91}$	$\mathbf{-16.01}$
tVAE	-23.6	-101.28	-88.75
JMVAE-Zero	-24.19	-131.05	-99.71

D. Camera & LiDAR Experiment

This experiment finally shows the ability of the M^{2}VAE to interact with the proposed re-encoding loss on simulated sensory observations. Training VAEs in general can become unstable on complex observations and we observed that the M^{2}VAE is even more susceptible to this behavior due to varying input dimensionalities. One can easily introduce more factors to balance the reconstruction losses, but this would not scale well in a hyper parameter search. Therefore, we first trained well elaborated β-VAEs on all modalities exclusively and use their encoders as a preprocessing to the M^{2}VAE. This allows the dimensionality reduction of all high dimensional modality inputs ($\sim 10^6$ for a camera vs. $\sim 10^3$ for LiDAR a frame) to a common size in dimension $D_\mathrm{a} = D_\mathrm{b} = \ldots$ which makes the introduction of further hyper parameters to the M^{2}VAE unnecessary.

We compare the setups using the standard scikit MLP classifier setup with and without M^{2}VAE embeddings. We performed the first classification task with a meta-sensor setup, which is able to sense all attributes at once, resulting in a classification rate of $96.2\,\%$ with M^{2}VAE versus $97.1\,\%$ without. We attribute the slight classification drawback to the generalization of VAEs, but take the results as a proof of concept for our M^{2}VAE approach, that raw – by means of β-VAE – and embedded – by means of β-VAE $\rightarrow$ M^{2}VAE – observations perform equally well.

Next, we compare the M^{2}VAE in the single-sensor setup without – resulting in $65.8\,\%$ – and with – resulting in $94.3\,\%$ – additional re-encoding loss during the training phase. The sensings are performed each with a single-sensor setup which is only able to perform consecutive sensing of attributes and, therefore, needs to facilitate in-place fusion to combine the sensings. The high performance impact shows the necessity of the proposed re-encoding loss, that almost reaches the desired performance which we, again, attribute to the generalization.

The M^{2}VAE was configured wrt. Sec. V-A with $D_* = 10$.

VI. Conclusion

We introduced the novel Variational Autoencoder framework M^{2}VAE for multi-modal data and showed that it can be trained on a variety of data sets. Furthermore, we developed the concept of in-place sensor fusion, which is applicable in distributed sensing scenarios and formulated its requirements by means of auto re-encoding. However, our introduced objective via the re-encoding loss facilitates in-place fusion and prevents label switching even on complex observations by maintaining the latent embedding during training. We performed all qualitative evaluations of the latent space with the premise in mind that a good generative model should not just generate good data but also gives a good latent representation, which correlates with the quantitative results. The M^{2}VAE is publicly available while future work will elaborate on the epistemic sensing and modality exchange.

References

[1] W. Wang, X. Yan, H. Lee, and K. Livescu, "Deep Variational Canonical Correlation Analysis," vol. 1, 2016. [Online]. Available: http://arxiv.org/abs/1610.03454

[2] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," *Proceedings of The 28th International Conference on Machine Learning (ICML)*, 2011.

[3] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," pp. 1–12, 2017.

[4] R. Vedantam, I. Fischer, J. Huang, and K. Murphy, "Generative Models of Visually Grounded Imagination," pp. 1–21, 2017. [Online]. Available: http://arxiv.org/abs/1705.10762

[5] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, and G. Deepmind, "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," *Iclr*, no. July, pp. 1–13, 2017.

[6] I. Higgins, A. Pal, A. A. Rusu, L. Matthey, C. P. Burgess, A. Pritzel, M. Botvinick, C. Blundell, and A. Lerchner, "DARLA: Improving Zero-Shot Transfer in Reinforcement Learning," 2017. [Online]. Available: http://arxiv.org/abs/1707.08475

[7] G. Liu, A. Siravuru, S. Prabhakar, M. M. Veloso, and G. Kantor, "Learning end-to-end multimodal sensor policies for autonomous navigation," *CoRR*, vol. abs/1705.10422, 2017. [Online]. Available: http://arxiv.org/abs/1705.10422

[8] T. Korthals, M. Kragh, P. Christiansen, H. Karstoft, R. N. Jørgensen, and U. Rückert, "Obstacle Detection and Mapping in Agriculture for Process Evaluation," *Frontiers in Robotics and AI Robotic Control Systems*, vol. 1, no. 1, 2018.

[9] R. Weston, S. Cen, P. Newman, and I. Posner, "Probably Unknown - Deep Inverse Sensor Modelling Radar," Tech. Rep., 2018. [Online]. Available: https://arxiv.org/abs/1810.08151

[10] A. Elfes, "Dynamic control of robot perception using multi-property inference grids," 1992.

[11] S. Herbrechtsmeier, T. Korthals, T. Schöpping, and U. Rückert, "AMiRo: A Modular & Customizable Open-Source Mini Robot Platform," in *ICSTCC*, 2016.

[12] M. E. Liggins, D. L. Hall, and D. Llinas, *Handbook of multisensor data fusion*, 2001.

[13] T. Korthals, D. Rudolph, M. Hesse, and R. Ulrich, "Multi-Modal Generative Models for Learning Epistemic Active Sensing," in *IEEE International Conference on Robotics and Automation*, Montreal, Canada, 2019.

[14] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *CoRR*, vol. abs/1312.6, 2013. [Online]. Available: http://arxiv.org/abs/1312.6114

[15] K. Sohn, H. Lee, and X. Yan, "Learning Structured Output Representation using Deep Conditional Generative Models," in *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015.

[16] G. Pandey and A. Dukkipati, "Variational methods for conditional multimodal deep learning," *Proceedings of the International Joint Conference on Neural Networks*, vol. May, pp. 308–315, 2017.

[17] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An Empirical Evaluation of Deep Architectures on Problems with Many Factors of Variation," ser. ICML. ACM, 2007, pp. 473–480.

[18] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient Learning of Sparse Representations with an Energy-based Model," ser. NIPS'06. MIT Press, 2006.

[19] C. Cadena, A. Dick, and I. D. Reid, "Multi-modal Auto-Encoders as Joint Estimators for Robotics Scene Understanding," in *RSS XIII*. Cambridge: MIT Press, 2016.

[20] T. Korthals, "M^{2}VAE – Derivation of a Multi-Modal Variational Autoencoder Objective from the Marginal Joint Log-Likelihood," 2019. [Online]. Available: http://arxiv.org/abs/1903.07303

[21] I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, and A. Lerchner, "Early Visual Concept Learning with Unsupervised Deep Learning," 2016. [Online]. Available: http://arxiv.org/abs/1606.05579

[22] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," 2018. [Online]. Available: http://arxiv.org/abs/1804.03599

[23] A. Dosovitskiy and T. Brox, "Generating Images with Perceptual Similarity Metrics based on Deep Networks," 2016.

[24] G. Perry, E. T. Rolls, and S. M. Stringer, "Continuous transformation learning of translation invariant representations," *Experimental Brain Research*, vol. 204, no. 2, pp. 255–270, 2010.

[25] M. F. Kragh, P. Christiansen, M. S. Laursen, M. Larsen, K. A. Steen, O. Green, H. Karstoft, and R. N. Jørgensen, "FieldSAFE: Dataset for Obstacle Detection in Agriculture," *Sensors*, vol. 17, no. 11, 2017.

[26] S. Herbrechtsmeier, T. Korthals, T. Schöpping, and U. Rückert, "AMiRo: A modular & customizable open-source mini robot platform," in *ICSTCC 2016*, 2016.

[27] K. Friston, "The free-energy principle: A unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[28] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder Variational Autoencoders," no. Nips, 2016. [Online]. Available: http://arxiv.org/abs/1602.02282