Multimodal Variational Autoencoders for Sensor Fusion and Cross Generation

Matthieu Da Silva–Filarder
Renault Software Factory
Eurecom, France
dasilva@eurecom.fr

Andrea Ancora
Renault Software Factory
andrea.ancora@renault.com

Maurizio Filippone
Eurecom, France
filippone@eurecom.fr

Pietro Michiardi
Eurecom, France
michiardi@eurecom.fr

***Abstract*—The cognitive system of humans, which allows them to create representations of their surroundings exploiting multiple senses, has inspired several applications to mimic this remarkable property. The key for learning rich representations of data collected by multiple, diverse sensors, is to design generative models that can ingest multimodal inputs, and merge them in a common space. This enables to: i) obtain a coherent generation of samples for all modalities, ii) enable cross-sensor generation, by using available modalities to generate missing ones and iii) exploit synergy across modalities, to increase reconstruction quality.**

In this work, we study multimodal variational autoencoders, and propose new methods for learning a joint representation that can both improve synergy and enable cross generation of missing sensor data. We evaluate these approaches on well-established datasets as well as on a new dataset that involves multimodal object detection with three modalities. Our results shed light on the role of joint posterior modeling and training objectives, indicating that even simple and efficient heuristics enable both synergy and cross generation properties to coexist.

I. Introduction

Models that learn form multiple sensor modalities have flourished recently [1, 2, 3, 4], because they yield rich, abstract and general representations, mimicking the way humans portray the surrounding environment based on multiple senses.

Multimodal systems, which exploit the combination of multiple inputs, are expected to learn improved data representations, when compared to unimodal approaches. This is assessed by measuring their *synergy* [5], which is defined as a gain in the quality of the generative model when using multiple modalities, as opposed to using a single modality alone. In addition, recent approaches address the challenging task of generating one modality conditioned on a different input modality, enabling a vast array of applications in which missing (or costly) sensor modalities are the norm rather than the exception. For example, in the medical domain, recent works [3] attempt at generating MRI portraits of a patient, given as an input a clinical diagnosis in textual form. In the automotive industry, the development of multimodal generative models can compensate the presence of faulty sensors, and allow to construct a rich representation of the environment despite sensor failures [6].

Although the presence of multiple modalities has the potential to offer additional information, that models can exploit to build richer representations, the key challenges to address revolve around i) the methodology to fuse the multimodal sensor data to infer a joint multimodal representation, that exploits the available synergies and obtains a gain in terms of representation compared to a unimodal model, and ii) efficient approaches to enable the generation of missing sensor modalities conditioned on the ones that are available, also called *cross generation* in the following parts. Recently, such challenges have attracted an increasing interest in multimodal generative models, with several variants of joint latent approximation methods, as well as explicit or implicit techniques to define training objectives that can account for missing input modalities. Despite encouraging results of multimodal systems robust to missing modalities [7], as well as fervid debates [8], there is no clear understanding, yet, of which method is a sensible choice to address the aforementioned challenges of multimodal representation learning.

We aim to fill this gap, and understand what are the synergy and cross generation properties of the various joint approximation methods, as well as whether these two properties can be achieved simultaneously. We do so in light of a series of objective metrics that help us understand the impact of the design choices underlying multimodal generative models based on the variational autoencoder (VAE) [9] structure. A sensible way to characterize the behavior of multimodal VAE is i) to measure their ability to generate all modalities, by sampling from the prior on the latent variables, ii) to measure cross generation by conditioning the latent representation on available modalities, and generating the missing ones, and finally iii) to study the synergy in terms of generative performance gains that emerge from a joint latent representation, when compared to unimodal generative variational autoencoders.

Our contributions are as follows:

- We propose a general formulation of the joint approximate posterior over the latent representation and design a series of heuristics to produce an approximate, joint latent space that are both simple to interpret and to implement.
- We propose a novel method to tackle the problem of modality cross generation. Our approach admits a simple objective to learn the model parameters, and can be applied to any existing method. Furthermore, our approach is more efficient than previous methods from the literature, which are traditionally combinatorial in the number of input modalities.
- We define an experimental protocol to assess the properties of a variety of methods. Our comparative study

involves several datasets of increasing difficulty, and uses established performance metrics.

II. Related Work

Variational generative models [9, 10] have attracted a lot of attention from the literature. Recent extensions focus on multiple modalities that share common information. The joint multimodal variational autoencoder (JMVAE) [11] marks a first attempt at modeling explicitly a joint distribution over a set of common generative parameters using the variational autoencoder framework. One of the main intent of JMVAE is to approximate the joint posterior distribution on this parameters set even when modalities are missing. To this extent, JMVAE requires as many encoder networks as possible subset of modalities, hence becoming impractical as the number of modalities increases.

As a remedy to these issues, the multimodal variational autoencoder MVAE [1] admits a conditional independence assumption on each modality with respect to the set of generative parameter. This assumption allows for a factorization of the joint posterior distribution over the set of generative parameters into a product of the unimodal posterior distributions over the same set. This results in a model that only requires as many encoders as input modalities and that can adapt to missing modalities configurations. While the MVAE methodology is applied in various fields [7, 12], it is found to suffer from low performance in generating missing modalities.

Following works [2, 4] interpret the factorization derivation of MVAE called "Product-of-experts" PoE to be the origin of the incapacity to generate missing modalities. They propose to use a "Mixture-of-experts" MoE combination, which translates into approximating the joint distribution approximation using an average of the unimodal distributions, hence replacing the product operation by a sum. Although outperforming competitors on several benchmarks, the mixture-of-experts variational autoencoder (MMVAE) fails to capture the benefits of the synergy among modalities.

Additional methods using variations of the PoE [13, 5] and MoE [14] approaches, aggregate multiple modalities in variational autoencoders but they do not address the aforementioned limitations of either methods. Other works [3, 15] address explicitly cross generation. For instance [15] proposes to randomly drop modalities during training, and [3] add loss terms to account for each one-to-one modality cross generation with many similarities to MMVAE. The recent work of [16] combines the PoE and MoE methods in a single model. While this method is general, it can prove difficult to scale to many modalities, due to the need for 2^M loss terms, with M the number of modalities.

In this paper, we present simple and efficient methods that overcome all the issues affecting prior works, while achieving comparable and often superior performance.

III. Method

Our multimodal generative model extends the VAE method [9, 10] to M modalities. Let $\mathbf{x}_i \in \mathbb{R}^{d_i}$, with $i \in \{1, \cdots, M\}$ denote observed variables describing the same phenomenon across M different modalities. Note that each observed modality may have different dimensions d_i. We use latent variables $\mathbf{z} \in \mathbb{R}^L$ to define a generative model over the joint distribution $p(\mathbf{x}_1, \cdots, \mathbf{x}_M, \mathbf{z}) = p(\mathbf{x}_{[1:M]}, \mathbf{z})$. By assuming conditional independence of the observed variables given the latent variable [1], i.e ., $\mathbf{x}_i \perp \mathbf{x}_j | \mathbf{z}$, $\forall i \neq j$, the joint distribution factorizes as:

$$p_{\boldsymbol{\Theta}}(\mathbf{x}_1, \cdots, \mathbf{x}_M, \mathbf{z}) = p(\mathbf{z}) \prod_{i=1}^{M} p_{\boldsymbol{\theta}_i}(\mathbf{x}_i|\mathbf{z}), \tag{1}$$

where $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the prior on the latent variables, which we assume to be an isotropic Gaussian distribution, and $p_{\boldsymbol{\theta}_i}(\mathbf{x}_i|\mathbf{z})$ are the likelihoods parametrized by deep neural networks, with parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}$. Our objective is to maximize the marginal likelihood of the data with respect to the latent variables $\mathbf{z}$:

$$p_{\boldsymbol{\Theta}}(\mathbf{x}_{[1:M]}) = \int p_{\boldsymbol{\Theta}}(\mathbf{x}_{[1:M]}, \mathbf{z}) d\mathbf{z} \tag{2}$$

$$= \int p(\mathbf{z}) \prod_{i=1}^{M} p_{\boldsymbol{\theta}_i}(\mathbf{x}_i|\mathbf{z}) d\mathbf{z}. \tag{3}$$

Unfortunately, computing the evidence is intractable, as it requires knowledge of the true joint posterior distribution $p(\mathbf{z}|\mathbf{x}_{[1:M]})$ that is unknown. Then, we approximate the true joint posterior with a variational joint posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_{[1:M]})$, and compute a lower bound to the marginal log likelihood, called the ELBO, as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{x}_{[1:M]}) =& \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_{[1:M]})} \left[\sum_{i=1}^{M} \log p_{\boldsymbol{\theta}_i}(\mathbf{x}_i|\mathbf{z}) \right] \\ & - \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_{[1:M]}) || p(\mathbf{z})),\end{aligned} \tag{4}$$

where the first term corresponds to a model fitting term, with a likelihood defined according to the modeling assumptions for a given task, and the second term is a regularization term that makes sure the approximate joint posterior does not deviate too much from the prior on the latent variables. In this work we assume the variational approximate joint posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_{[1:M]}) \in \mathbf{Q}$ to be a "simple" distribution that is easy to sample from (e.g., a Gaussian distribution).

A. Joint Approximate Posterior

The joint approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_{[1:M]})$ is the key to learn a useful mapping from observed to latent variables. The approximation of the joint posterior should benefit from the addition of new modalities, either by the addition of new information, or in improved reliability where information is common with other modalities. Joint and cross generation should be coherent across modalities, which calls for a training method that allows for missing modalities [1, 2]. A joint generative model should exploit the synergy across all modalities, and produce tangible improvements when compared to unimodal generative models [2, 5].

In this work we introduce three joint approximation methods: the first can be thought of as a generalization of the PoE [17],

whereas the other two are heuristics that have the merit of being simple and computationally efficient.

Robust Bayesian Committee Machine. We use the conditional independence assumption in the generative model to derive a relation among joint- and single-modality posteriors through the *robust Bayesian committee machine* (RBCM) joint posterior [18, 19]:

$$p^{\text{RBCM}}(\mathbf{z}|\mathbf{x}_{1:M}) \propto \frac{\prod_{i=1}^{M} p^{\beta_i}(\mathbf{z}|\mathbf{x}_i)}{p^{-1+\sum_i \beta_i}(\mathbf{z})}, \tag{5}$$

where the coefficients β_i add the flexibility of increasing/reducing the importance of experts, and are computed as the difference in the differential entropy between the prior and the unimodal posterior. If the true posteriors for each individual factor $p(\mathbf{z}|\mathbf{x}_i)$ is properly contained in the family of its variational counterpart, $q_{\boldsymbol{\phi}_i}(\mathbf{z}|\mathbf{x}_i)$, then we can write:

$$p^{\text{RBCM}}(\mathbf{z}|\mathbf{x}_{1:M}) \approx q_{\boldsymbol{\phi}}^{\text{RBCM}}(\mathbf{z}|\mathbf{x}_{1:M}) = \frac{\prod_{i=1}^{M} q_{\boldsymbol{\phi}_i}^{\beta_i}(\mathbf{z}|\mathbf{x}_i)}{p^{-1+\sum_i \beta_i}(\mathbf{z})}. \tag{6}$$

In this work, we assume the unimodal approximate posteriors to be Gaussian distributed with diagonal covariance matrix, that is $q_{\boldsymbol{\phi}_i}(\mathbf{z}|\mathbf{x}_i) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i(\mathbf{x}_i), \boldsymbol{\sigma}_i^2(\mathbf{x}_i)\mathbf{I})$ where i denotes a modality index. Then, by eq. (6), the joint approximate posterior is also Gaussian with diagonal covariance matrix, $q_{\boldsymbol{\phi}}^{\text{RBCM}}(\mathbf{z}|\mathbf{x}_{1:M}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}^{\text{RBCM}}(\mathbf{x}_{[1:M]}), (\boldsymbol{\sigma}^{\text{RBCM}})^2(\mathbf{x}_{[1:M]})\mathbf{I}\right)$, with:

$$\boldsymbol{\mu}^{\text{RBCM}} = (\boldsymbol{\sigma}^{\text{RBCM}})^2 \sum_{i=1}^{M} \beta_i \boldsymbol{\sigma}_i^{-2} \boldsymbol{\mu}_i, \tag{7}$$

$$(\boldsymbol{\sigma}^{\text{RBCM}})^{-2} = \sum_{i=1}^{M} \beta_i \boldsymbol{\sigma}_i^{-2} + \left(1 - \sum_{i=1}^{M} \beta_i\right) \boldsymbol{\sigma}_p^{-2}, \tag{8}$$

where $\boldsymbol{\sigma}_p$ is the prior variance. Our formulation combines the flexibility of the generalized POE with an appropriate Bayesian treatment of the joint posterior. This induces the term $\left(1 - \sum_{i=1}^{M} \beta_i\right)$ in the posterior precision, which ensures a proper fallback to the prior. The coefficients β_i control the importance of unimodal approximate posteriors (individual experts), and how strong the influence of the prior is. As done in [20], for computational efficiency, we use $\beta_i = \frac{1}{2}(\log \boldsymbol{\sigma}_p^2 - \log \boldsymbol{\sigma}_i^2(\mathbf{x}_i))$. In the often used hypothesis of a Gaussian posterior with diagonal covariance, we can expand the RBCM method to the latent dimension level using the independence assumption:

$$q_{\boldsymbol{\phi}_i}^{\boldsymbol{\beta}_i}(\mathbf{z}|\mathbf{x}_i) = \prod_{j=1}^{L} q_{\boldsymbol{\phi}_i}^{\beta_{i,j}}(\mathbf{z}_j|\mathbf{x}_i) \tag{9}$$

where z_j is the element at the j^{th} dimension of the latent vector $\mathbf{z}$, and $\boldsymbol{\beta}_i = (\beta_{i,1}, ..., \beta_{i,L})$ that is now a vector composed by $\beta_{i,j} = \frac{1}{2}(\log \sigma_p^2 - \log \sigma_{i,j}^2(\mathbf{x}_i))$ with $\sigma_{i,j}$ the j^{th} element of $\boldsymbol{\sigma}_i$. This expansion allows for a fine-grained control of the importance of any unimodal approximate posterior when compared to the prior at a certain dimension of the latent.

Joint posterior heuristic approximations. Instead of searching for complex aggregation schemes, we consider simple approaches that trade approximation quality for improved computational efficiency. We build on the key intuition of RBCM, which gauges the mixing of unimodal components to produce the joint approximation, and derive two heuristics where only individual unimodal components contribute to the joint approximation. The first, called *best component expert* (BCE), approximates the joint posterior by choosing the unimodal approximate posterior that has the lowest variance for each element of the joint latent space. In the unimodal Gaussian distribution setting with diagonal covariance matrix, we have that $q_{\boldsymbol{\phi}}^{\text{BCE}}(\mathbf{z}|\mathbf{x}_{1:M}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}^{\text{BCE}}(\mathbf{x}_{[1:M]}), (\boldsymbol{\sigma}^{\text{BCE}})^2(\mathbf{x}_{[1:M]})\mathbf{I}\right)$, where

$$\mu_j^{\text{BCE}} = \mu_{i^*,j} \text{ with } i^* = \underset{i \in [1,M]}{\text{argmin}}\{\sigma_{i,j}\}, \forall j \in [1, L] \tag{10}$$

$$\sigma_j^{\text{BCE}} = \min_{i \in [1,M]} \{\sigma_{i,j}\}, \forall j \in [1, L] \tag{11}$$

where $\sigma_{i,j}$ is the j^{th} element of the variance $\boldsymbol{\sigma}_i$ associated to modality i, and L is the size of the latent space. This heuristic can be interpreted as a special case of RBCM where $\beta_{i,j} = 1$ if $i = \underset{i}{\text{argmin}}(\sigma_{i,j})$ and $\beta_{i,j} = 0$ otherwise.

The second heuristic we introduce uses randomization. Instead of choosing the best unimodal approximate posterior as done with the BCE heuristic, we choose the unimodal elements at random among the modalities, without taking into consideration any confidence indicator. This heuristic is called *random component expert* (RCE). Under the same assumptions for the unimodal approximate posterior distributions we used so far, the resulting RCE joint posterior approximation is also a Gaussian distribution with diagonal covariance matrix $q_{\boldsymbol{\phi}}^{\text{RCE}}(\mathbf{z}|\mathbf{x}_{1:M}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}^{\text{RCE}}(\mathbf{x}_{[1:M]}), (\boldsymbol{\sigma}^{\text{RCE}})^2(\mathbf{x}_{[1:M]})\mathbf{I}\right)$. Each dimension of $\boldsymbol{\mu}^{\text{RCE}}(\mathbf{x}_{[1:M]})$ and $\boldsymbol{\sigma}^{\text{RCE}}(\mathbf{x}_{[1:M]})$ is independently and uniformly sampled from the set of approximate unimodal means and standard deviations at the same index. The RCE heuristic can be seen as a generalization of the MOE presented in [2] that selects instead a unique expert to contribute to the joint posterior, for all latent dimensions.

B. Missing Modalities and Cross Generation

A key property of multimodal generative models is their ability to work with missing modalities and to enable cross generation, that is to use a subset of input modalities to generate output modalities missing from the input subset. The ELBO formulation in eq. (4) does not explicitly consider such requirements. The sub-sampling training paradigm [1], explicitly modifies the ELBO by introducing unimodal and multimodal terms, but fails to account for the generation of missing modalities. The MOE method [2] implicitly defines an ELBO that accounts for cross generation, but prevents the synergy between modalities in the latent space.

A novel approach to enable both cross generation and synergy is thus truly needed. We introduce two new methods.

Exhaustive cross generation. We modify the ELBO in eq. (4) by introducing terms corresponding to all possible subsets of inputs of modalities. We label the new ELBO *exhaustive cross generation* (ECG), and define it as follows:

$$\begin{aligned}\mathcal{L}_{\text{ECG}}(\mathbf{x}_{[1:M]}) &= \mathcal{L}(\mathbf{x}_{[1:M]}) \\ &+ \sum_{j=1}^{2^M-1} \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|X_j)} \left[\sum_{i=1}^{M} \log p_{\boldsymbol{\theta}_i}(\mathbf{x}_i|\mathbf{z})\right] \\ &- \text{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|X_j)||p(\mathbf{z})\right)\end{aligned} \tag{12}$$

with each X_j being a different subset of $\mathbf{x}_{[1:M]}$ with at least one modality. This "brute-force" approach is computationally expensive, as it requires $2^M - 1$ forward passes to account for each possible input modality set, but establishes an upper bound on cross generation quality.

An alternative is to take into account only a subset of all possible subsets to reduce computational requirements. This method echoes with some ideas developed by [15] and with the sub-sampling paradigm presented in MVAE [1] but with the crucial addition of penalizing also the reconstruction of the missing modalities.

Latent component dropout. The novel approach we introduce is based on randomization, and draws inspiration from the information dropout method [21]. The gist of the idea is to "simulate" missing modalities by inducing randomization through a dropout mask. We apply a dropout mask to individual elements of the latent variables, one for each modality, and revert to a prior expert for those elements selected by the mask. We call our method *latent component dropout* (LCD), and apply it to the joint approximate posterior.

In detail, we define the dropout mask to be a vector of Bernoulli distributed random variables: $\mathbf{m}_i \sim \mathcal{B}(1-\alpha)$, $\forall i \in \{1, \cdots, M\}$ with $\mathbf{m}_i \in \{0,1\}^L$. When we apply $\mathbf{m}_i$ to the i-th approximate unimodal posterior, we obtain $q^{\text{LCD}}_{\boldsymbol{\phi}_i}(\mathbf{z}|\mathbf{x}_i) = \mathcal{N}(\mathbf{z}; \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i^2\mathbf{I})$, with moments defined as:

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i \odot \mathbf{m}_i + \boldsymbol{\mu}_p \odot (1 - \mathbf{m}_i), \tag{13}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \boldsymbol{\sigma}_i^2 \odot \mathbf{m}_i + \boldsymbol{\sigma}_p^2 \odot (1 - \mathbf{m}_i), \tag{14}$$

where $\boldsymbol{\mu}_p$ and $\boldsymbol{\sigma}_p^2$ are the prior mean and variance, and $\odot$ is the Hadamard product. Our method induces a fine-grained cross generation, whereby means and variances of individual elements of the latent variables result from a randomized subset of those elements across all modalities. Latent component dropout, combined with any joint approximate posterior, is computationally efficient. Furthermore, it alleviates the problems of the MoE approach that requires M^2 ELBO terms in the training loss, while admitting any joint approximation method that can aggregate multiple unimodal approximations.

IV. Experiments

We study the impact of the joint posterior formulation on metrics related to multi-modality, and which variant of the variational objective better satisfies cross generation requirements. We do so by comparing several methods from the literature to our proposed approaches, using several datasets and various metrics that measure important criteria related to multi-modality. The methods we study include: MVAE [1], MMVAE [2], and JMVAE [11]. We compare three cross generation methods: exhaustive cross generation (ECG) and latent component dropout (LCD) that we propose, and random modality dropout (RMD) [15], which uses a random subset of modalities at each training step.

For our experiments we use Gaussian priors and approximate posteriors, we use the Adam optimizer [22] with a learning rate of 0.001 for MNIST-SVHN and 0.0005 for the other datasets. All results are reported over 5 runs with different seeds.

A. Datasets

MNIST-SVHN. This is a simple and widely used multimodal dataset [2, 5]. We train our models using pairs of MNIST and street view house numbers (SVHN) images, which share the same class: each MNIST sample is randomly paired with an instance of the same class from SVHN. We use a convolutional neural network (CNN) architecture for SVHN, and a fully connected neural network for MNIST. We align training and testing splits on MNIST (60000 training and 10000 testing samples).

Fig. 1: MNIST-SVHN examples of matching modalities.

CUB Image-Captions. This is a challenging language-vision multimodal dataset, also used in [2]. We consider 11,788 photos of birds, annotated with 10 fine-grained captions describing the bird's appearance. We use 2048 dimensional image feature vectors extracted using a pre-trained ResNet-101 [23] and 300 dimensional FastText [24] embedding vectors for the text modality. The encoder and decoder are fully connected neural networks for the image modality and CNNs for the embedded text modality. The train/test split for this dataset is 75%/25%.

Caption: This white bird has black along the ends of its wings and a pale, long beak.

Fig. 2: CUB Image-Captions example image and caption.

Multimodal Shapes 3D. We introduce a new dataset inspired from 3D shapes [25] with 3 modalities: image, depth map and a radar-like modality. This dataset of 17,729 scenes was generated using Blender [26], and contains scenes with

multiple objects, and their associated bounding boxes obtained from the depth map. The radar modality is generated as a post-processing of the depth maps with ground points removal. We use CNN architectures for encoding and decoding the three modalities. The train/test split for this dataset is 80%/20%.

B. Metrics

We revisit metrics introduced in [2, 5]. These metrics are based on the shared information between the various modalities: digit label for MNIST-SVHN, caption matching the bird picture for CUB IMAGE-CAPTIONS, and object types and positions in the scenes for MULTIMODAL SHAPES 3D. For the MNIST-SVHN dataset the shared information is extracted with pre-trained digit classifiers while for the MULTIMODAL SHAPES 3D dataset we use a pre-trained single-shot object detector for each modality. On the CUB IMAGE-CAPTIONS dataset we use the similarity metric proposed in [2] directly on the caption and image. Next, to keep the notation fluid, the shared information will be referred as "label".

Joint fractional agreement. By generating samples for all modalities from the prior, we measure the fraction of times the labels predicted by a pre-trained classifier on the generated modalities are in agreement with each other.

Cross fractional agreement. By generating samples for a given missing modality, conditioned on the available modalities, we measure the similarity between ground truth and predicted labels associated to the available modalities. When there are more than 2 modalities, the missing modality is reconstructed using all other input modalities.

Synergy. Our goal is to quantify the benefits of a multimodal representation, compared to a unimodal VAE. Thus, we report the gains obtained by using all modalities compared to an individual VAE for each modality. The gain $\mathcal{G}$ is computed by $\mathcal{G} = \frac{S-U}{U}$, where S is the synergy score and U the unimodal VAE score. For MMVAE, we report results based on the best modality as opposed to a random one, to overcome the limitations of the MOE approach, and have a fair evaluation. Note also, that for synergy and cross fractional agreement we have one score for each modality.

C. Results

MNIST-SVHN: We present quantitative results using the common digit as the shared information between the two modalities. We use for all models the same encoders and decoders architecture except for JMVAE that requires additional encoders and decoders. For all methods the latent space size is $L = 20$, we use gaussian priors, and all methods use importance weighted auto-encoder [27] with the number of importance samples set to $K = 30$.

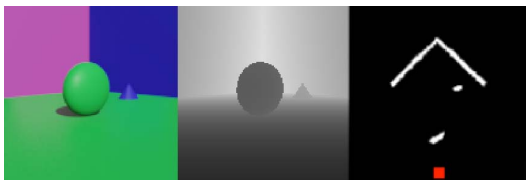


Fig. 3: MULTIMODAL SHAPES 3D example. On the radar modality the red square denotes the camera position.

In Tab. I we report joint, cross and synergy metrics. Concerning the joint metric, it is striking to notice that the choice of a joint approximate posterior method has negligible effects on performance, especially when cross generation is omitted. Instead, the ECG cross generation method has a positive impact on joint metric, which is surprising as it does not directly optimize for joint modality generation. Concerning cross fractional agreement, note that the task of reconstructing a digit from one modality to the other is more challenging from SVHN to MNIST because the digits from SVHN are harder to recognize. In general, by a thorough inspection of the latent spaces, we notice that when no cross generation method is used during training, the models we evaluate implicitly prefer one direction to the other, and exhibit inconsistent performance.

Synergy results in Tab. I compare the digit classification accuracy of the multimodal VAE architectures when all modalities are present against two unimodal VAE models reconstructing independently the two modalities. The mean digit classification scores for unimodal VAEs reconstructions are 92.9% for MNIST and 36.2% for SVHN. The synergy gains of a multimodal approach are low on the MNIST modality because there is little margin of improvement, since it is an "easy" classification problem. Instead, on the SVHN modality, synergy gains are high for all methods, with gains up to $+142.7\%$ for BCE with exhaustive cross generation. As expected, MMVAE is the method struggling the most on the synergy metric due to the MOE properties.

In the last two rows of Tab. I we compare RMD and LCD cross generation methods while fixing the joint approximation to BCE. Despite not reaching results as high as ECG, the cross and synergy scores of LCD are competitive with RMD while only requiring a single ELBO term. It is interesting to notice that our proposed joint posterior approximation method RCE, which implicitly contributes to cross generation, scores similarly to models that employ an explicit ELBO for cross generation. This indicates, that our randomized posterior approximation method is a good, and simple candidate to be used if the computational budget is limited.

CUB IMAGE-CAPTIONS. For this dataset, to compute our metrics, we use the canonical correlation analysis (CCA) [2] between image and text embeddings computed using a pre-trained Resnet-101 and FastText [24] models. The same encoders, decoders, latent size of $L = 64$ and number of importance weighted auto-encoder samples of $K = 10$ are used for all experiments except for JMVAE. Results are depicted in Tab. II, where we focus on cross generation and synergy, since these are the tasks that have been overlooked in the literature, and because they are relevant from an application point of view (e.g., the automotive domain).

Results on cross fractional agreement highlight the importance of cross generation techniques like ECG, RMD and LCD as they significantly outperform methods without cross

Cross Gen. method	Joint Approx. method	Joint Frac. Agr.	Cross Frac. Agr.		Synergy Gain (%)	
			S ->M	**M ->S**	**all ->M**	**all ->S**
Implicit	JMVAE	52.57 (±2.15)	**73.83** (±1.39)	59.21 (±9.37)	−1.3	+126.9
	MMVAE	47.02 (±2.58)	54.42 (±5.12)	77.81 (±4.47)	−2.0	+90.7
	RCE *	49.43 (±2.90)	67.00 (±5.71)	74.59 (±2.61)	−1.4	+133.9
None	MVAE	52.92 (±1.16)	15.91 (±0.68)	68.06 (±9.67)	−0.4	+126.0
	BCE *	53.98 (±0.37)	10.11 (±0.21)	*87.06* (±1.08)	+0.0	+123.5
	RBCM *	52.47 (±0.38)	15.26 (±2.48)	57.71 (±15.31)	−0.2	+122.5
ECG *	MVAE	**58.91** (±1.05)	**73.77** (±1.60)	**81.90** (±1.15)	**+1.5**	**+141.7**
	BCE *	**58.87** (±1.26)	**73.50** (±2.78)	77.27 (±3.26)	**+1.5**	**+142.7**
	RBCM *	**58.82** (±1.31)	70.38 (±2.59)	**81.81** (±1.78)	**+1.0**	**+141.5**
RMD	BCE *	55.44 (±1.37)	63.00 (±2.57)	67.28 (±9.82)	+0.5	+136.4
LCD *	BCE *	52.96 (±1.73)	70.04 (±1.85)	65.32 (±8.37)	+0.0	+138.3

TABLE I: Digit classification accuracy (%) on joint, cross fractional agreement and synergy. Our methods are marked with (*).

Cross Gen. method	Joint Approx. method	Cross Frac. Agr.		Synergy Gain (%)	
		C ->I	**I ->C**	**all->I**	**all->C**
Implicit	JMVAE	0.14 (±0.01)	0.04 (±0.03)	−50.2	+32.3
	MMVAE	0.17 (±0.01)	0.04 (±0.01)	−39.4	+21.0
	RCE *	**0.45** (±0.01)	**0.17** (±0.01)	**+27.5**	**+214.2**
None	MVAE	−0.01 (±0.01)	0.09 (±0.01)	+1.8	+80.7
	BCE *	0.03 (±0.01)	0.08 (±0.02)	+4.6	+103.2
	RBCM *	0.01 (±0.01)	0.09 (±0.01)	+3.4	+93.5
ECG *	MVAE	**0.44** (±0.02)	**0.15** (±0.03)	+24.7	+163.6
	BCE *	**0.47** (±0.01)	0.10 (±0.02)	**+36.5**	+137.3
	RBCM *	0.40 (±0.02)	0.11 (±0.02)	+21.8	+121.8
RMD	BCE *	0.40 (±0.02)	**0.12** (±0.01)	**+27.5**	**+171.1**
LCD *	BCE *	**0.44** (±0.01)	0.08 (±0.00)	**+34.1**	**+174.2**

TABLE II: CCA coefficients on CUB IMAGE-CAPTIONS for cross fractional agreement and synergy gain, compared to an image-only VAE with a CCA of 0.276 and a caption-only VAE with a CCA of 0.062.

generation and state of the art methods for cross generation such as JMVAE and MMVAE. It is worth noticing on this dataset that RCE is as efficient as other cross generation techniques. In addition, a clear gain in synergy is apparent for all methods except for MMVAE due to the MOE formulation, and, more surprisingly for JMVAE, probably due to the more challenging nature of this dataset combined to the complexity of JMVAE architectures. As the CCA metric between image and caption puts an emphasis on the match between reconstructed modalities and the other modality ground truth, synergy gains are best for models trained with cross generation techniques compared no cross generation. The gap in synergy gain also denotes the improvement in representation quality gained through the use of cross generation technique. As for the comparison between cross generation techniques, LCD is performing slightly better than RMD and achieves results comparable to the exhaustive (thus costly) ECG method, even surpassing it on the synergy reconstruction of caption. Similarly to what is discussed for the MNIST-SVHN dataset, our proposed RCE method appears to be a viable and simple choice, in that it achieves competitive performance both in terms of cross generation and synergy.

MULTIMODAL SHAPES 3D: The three modalities of this dataset share the information of the objects type and position in the 3D scenes. We pre-train the the YOLOv3 [28] object detector on images and depth maps to detect the objects and infer their class along with their 2D bounding boxes. To evaluate the image and depth map reconstruction quality we compute a mean Average Precision (mAP) between the bounding box of the shapes detected on a reconstructed modality and their ground truth positions. We use this score to evaluate the synergy and cross fractional agreement metrics on the image modality by measuring the mAP scores obtained on the image reconstructions. The image, depth map, and radar samples are encoded and decoded by 2D CNNs models. We use the same encoders, decoders, latent size $L = 256$ for all models, and use classical VAE instead of importance weighted VAE. Also, we omit JMVAE, due to its poor scalability, since it requires 7 different encoders to cover for all possible input modality configurations.

Cross Gen. method	Joint Approx. method	Cross Frac. Agr.		Synergy Gain (%)	
		D+R ->I	**I+R ->D**	**all->I**	**all->D**
Implicit	MMVAE	0.51 (±0.02)	0.52 (±0.07)	−32.6	−40.2
	RCE *	0.93 (±0.01)	0.88 (±0.01)	+3.7	+2.5
None	MVAE	0.70 (±0.03)	0.01 (±0.01)	**+5.6**	**+6.7**
	BCE *	0.76 (±0.01)	0.02 (±0.01)	**+5.6**	**+6.4**
	RBCM *	0.75 (±0.01)	0.01 (±0.00)	**+5.6**	**+6.6**
ECG *	MVAE	**0.95** (±0.00)	**0.91** (±0.00)	**+5.3**	**+6.1**
	BCE *	0.93 (±0.01)	0.87 (±0.01)	+4.0	+3.4
	RBCM *	**0.95** (±0.00)	**0.91** (±0.01)	+5.2	+5.9
LCD *	MVAE	**0.94** (±0.00)	**0.90** (±0.01)	+4.6	+4.5
	BCE *	0.93 (±0.01)	**0.90** (±0.01)	+4.3	+4.5
	RBCM *	**0.94** (±0.01)	**0.90** (±0.00)	+4.7	+4.8

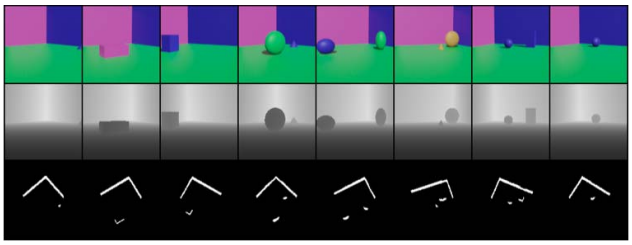
TABLE III: Object detection scores (mAP) on MULTIMODAL SHAPES 3D for cross fractional agreement and synergy (D: depth, I: image, R: radar). Unimodal VAE performance is: depth VAE mAP= 0.87, image VAE mAP= 0.91.

Comparing the cross fractional agreement results on the MULTIMODAL SHAPES 3D dataset presented in Tab. III we notice that methods without cross generation are unsuccessful at generating a coherent depth modality from image and radar. On the other hand there is a significant performance increase when using a cross generation, both implicit or explicit, but not for MMVAE. Indeed, MMVAE is underperforming compared to previous experiments because two modalities (depth and radar) are available to reconstruct the image modality, while MOE fusion can only exploit a single modality randomly sampled from the two available ones. The ECG method is the best performing on cross fractional agreement although it requires

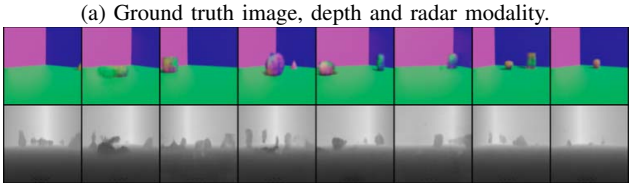
the computation of 7 loss terms at each training iteration. Efficient methods such as LCD or RCE are performing almost on par with ECG, but are more computationally efficient.

On the synergy gain, we notice improvements for all methods except for MMVAE that, as stated before, only exploits a single randomly chosen modality. Methods trained without cross generation or with ECG can reach up to $+6.7\%$ synergy gain for "None" cross generation method with MVAE, which corresponds to a mAP of 0.93 for the depth modality. The slightly better synergy gains from using no cross generation method can be explained by the "None" cross generation method being entirely focused on synergy with a single ELBO term taking into account the entire set of modalities. Instead, ECG has 7 ELBO terms to account for missing modality cases with only a single ELBO corresponding to the case when all modalities are present. As a result, the contribution of the single ELBO term related to synergy score of ECG is slightly weakened in the global objective. This minor loss in synergy compared to "None" cross generation is compensated by the substantial improvement gained on cross fractional agreement.

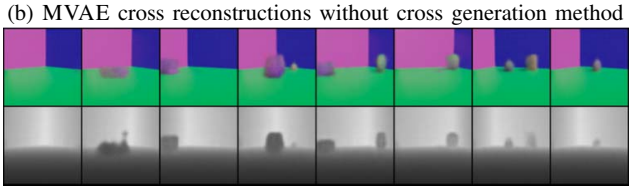
We show qualitative results in Fig. 4 with samples of the image modality generated from depth and radar, and the depth modality generated from image and radar. We notice the weakness of MVAE without cross generation that cannot obtain a coherent reconstruction from image and radar modalities to depth. For MMVAE, we remark that the object position is maintained, but shapes are often blurry and difficult to identify. The quality of the reconstructions from RBCM and RCE methods is comparable, with clearly defined shapes in both generated modalities. It is interesting to notice how images generated from depth and radar modalities deal with object coloring, with MMVAE, for instance, coloring objects in dull colors and MVAE with no cross generation often coloring the shapes with more than a single color.



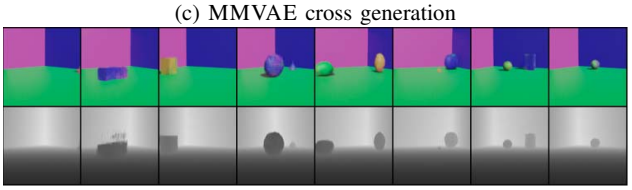
(a) Ground truth image, depth and radar modality.



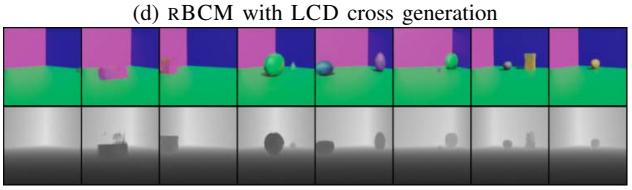
(b) MVAE cross reconstructions without cross generation method



(c) MMVAE cross generation



(d) RBCM with LCD cross generation



(e) RCE cross generation

Fig. 4: Cross generated samples on MULTIMODAL SHAPES 3D (D: depth, I: image, R: radar). (a) Ground true samples. (b)-(e) cross generation samples for various model configurations, top row: D+R $\rightarrow$ I, bottom row: I+R $\rightarrow$ D.

V. CONCLUSION

In this paper, we studied multimodal VAEs that are promising models for their application to sensor fusion in multi-sensor systems. Indeed, these models can exploit synergy between sensing data and are robust to missing or faulty sensors.

With the aim to improve over current multimodal VAEs models on synergy and cross generation properties, we proposed a range of candidate joint approximation and cross generation methods. In addition, to evaluate these methods, we introduced a new dataset that acts as a proxy to multimodal sensor systems applied to autonomous driving as it contains images, depth maps and radar-like modalities. Through extensive experiments on various multimodal datasets, we remarked that existing state-of-the-art methods often cannot reach satisfying results on both synergy and cross generation. In order to design multimodal systems that combine both such properties, we showed that even simple heuristic to approximate a joint latent representation can be a viable alternative to existing methods and their generalization. We also proposed three cross generation methods that are compatible with any approximate, joint latent representation methods. We assessed the trade-off between scalability and performance of these cross generation methods: in cases where the computational cost is not an issue, our proposed ECG method often provides the best results on both synergy and cross generation. However, when the number of input sensors grows, our proposed LCD heuristic achieves excellent performance while being extremely efficient. Finally, we also showed that our randomized heuristic RCE, that learns a joint latent space and caters to cross generation in a single, simple objective, was on par with other methods in a variety of settings.

The methods we proposed in this work open the doors to the application of multimodal VAEs to large multimodal systems, making them both efficient at merging information into a common latent representation, and resilient to faulty or missing

sensor inputs. The ability of our methods to generate missing modalities paves the way for sensor virtualization, whereby a sensor that has been used to train a model is intentionally removed or disabled (e.g., because it is costly), while being reconstructed through the other available modalities.

Our next steps concern empirical validation of our models using more realistic datasets from the automotive domain, which also call for additional work to include temporal sequences of sensor data. On the methodological side, we will explore disentanglement properties of the joint latent spaces, to enable interpretation of the learned representations, and the study of generalization properties of our models.

VI. Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation 2020-AD011012166 made by GENCI.

References

[1] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," 2018.

[2] Y. Shi, N. Siddharth, B. Paige, and P. H. S. Torr, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," 2019.

[3] L. Antelmi, N. Ayache, P. Robert, and M. Lorenzi, "Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data," in *International Conference on Machine Learning*. PMLR, 2019, pp. 302–311.

[4] R. Kurle, S. Günnemann, and P. Van der Smagt, "Multi-source neural variational inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4114–4121.

[5] Y. Shi, B. Paige, P. H. S. Torr, and N. Siddharth, "Relating by contrasting: A data-efficient framework for multimodal generative models," 2020.

[6] C. Cadena, A. R. Dick, and I. D. Reid, "Multi-modal auto-encoders as joint estimators for robotics scene understanding." in *Robotics: Science and Systems*, vol. 5, 2016, p. 1.

[7] R. Dorent, S. Joutard, M. Modat, S. Ourselin, and T. Vercauteren, "Hetero-modal variational encoder-decoder for joint modality completion and segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 74–82.

[8] S. Kutuzova, O. Krause, D. McCloskey, M. Nielsen, and C. Igel, "Multimodal variational autoencoders for semi-supervised learning: In defense of product-of-experts," 2021.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[10] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," 2014.

[11] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," 2016.

[12] M. Baruah and B. Banerjee, "A multimodal predictive agent model for human interaction generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1022–1023.

[13] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *International Conference on Learning Representations*, 2018.

[14] T. Sutter, I. Daunhawer, and J. E. Vogt, "Multimodal generative learning utilizing jensen-shannon divergence," in *Workshop on Visually Grounded Interaction and Language at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[15] M. Vasco, F. S. Melo, and A. Paiva, "Mhvae: a human-inspired deep hierarchical generative model for multimodal representation learning," *arXiv preprint arXiv:2006.02991*, 2020.

[16] T. M. Sutter, I. Daunhawer, and J. E. Vogt, "Generalized multimodal elbo," *arXiv preprint arXiv:2105.02470*, 2021.

[17] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[18] V. Tresp, "A bayesian committee machine," *Neural computation*, vol. 12, pp. 2719–41, 12 2000.

[19] M. P. Deisenroth and J. W. Ng, "Distributed gaussian processes," 2015.

[20] Y. Cao and D. J. Fleet, "Generalized product of experts for automatic and principled fusion of gaussian process predictions," 2015.

[21] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," 2017.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[25] C. Burgess and H. Kim, "3d shapes dataset," https://github.com/deepmind/3dshapes-dataset/, 2018.

[26] Blender Online Community, "Blender - a 3d modeling and rendering package," Blender Foundation, Blender Institute, Amsterdam, 2020. [Online]. Available: http://www.blender.org

[27] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," 2016.

[28] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.