Cross-modal Variational Alignment of Latent Spaces

Thomas Theodoridis Theocharis Chatzis Vassilios Solachidis Kosmas Dimitropoulos
Petros Daras
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece
{tomastheod, hatzis, vsol, dimitrop, daras}@iti.gr

Abstract

In this paper, we propose a novel cross-modal variational alignment method in order to process and relate information across different modalities. The proposed approach consists of two variational autoencoder (VAE) networks which generate and model the latent space of each modality. The first network is a multi-modal variational autoencoder that maps directly one modality to the other, while the second one is a single-modal variational autoencoder. In order to associate the two spaces, we apply variational alignment, which acts as a translation mechanism that projects the latent space of the first VAE onto the one of the single-modal VAE through an intermediate distribution. Experimental results on four well-known datasets, covering two different application domains (food image analysis and 3D hand pose estimation), show the generality of the proposed method and its superiority against a number of state-of-the-art approaches.

1. Introduction

Cross-modal learning has attracted increasing attention recently due to the rapid growth of multi-modal data (image, video, text, audio, depth, IR etc) and the need for enhanced learning either by leveraging information from one data modality to accomplish a given task in another, or through the synergistic synthesis of information from multiple modalities. Because of their general nature, they have been extensively used in the literature for various problems, such as audio retrieval from text [20], text-to-image and image-to-text retrieval [24], sentiment analysis from video, audio and text sources [22], synchronization among different representations of music, like sheet music and audio recordings [15], recipe (ingredients and instructions) retrieval from images and vice versa [27] and 3D hand pose estimation from images [30]. Recent cross-modal frameworks involve neural networks as encoder and decoder mechanisms in order to transition from one modality to another. Based on the way these frameworks model the cross-modal objective, they are categorized as discriminative and generative. Approaches that fall into the first category model the probability of an outcome conditioned on the given observation. Generative approaches, on the other hand, model the underlying distribution of the observed variables, thus obtaining valuable information regarding their origin.

Most recent approaches have adopted deep generative models, such as VAEs, GANs or a combination of them, to encode cross-modal data into a shared latent space [30, 34]. However, the main problem in these approaches is the fact that each modality has completely different characteristics from the others and, as a result, it is difficult to efficiently model the heterogeneous modalities (like image, speech or text) into a shared latent space. To address the problem of learning meaningful mappings among embedding spaces, we propose a novel variational alignment framework of latent spaces, which performs the mapping of the latent space of one modality onto the one of another modality. More specifically, in this paper we present a cross-modal learning approach consisting of a number of variational autonencoder networks that aim to generate and model the latent space corresponding to each modality and, at the same time, align the different spaces through the modeling of an intermediate latent space, generated by an additional variational autoneconder network. The main contributions of this paper are summarized as follows:

- We introduce a generic cross modal deep learning approach using variational autoendonder networks in order to model the latent spaces of different modalities as probability distributions. More specifically, we propose the use of a pair of multimodal (M_1-to-M_2) and single-modal (M_2-to-M_2) variational autoencoders with aligned latent spaces, where the aligned latent space of the first modality can be directly used by the decoder of the single-modal VAE network, outperforming the state-of-the-art in different application domains.

- We propose a novel cross-modal variational alignment

of the probability distributions of latent spaces corresponding to different modalities. By generating and modeling an intermediate latent space through a variational network, we can achieve better alignment of latent spaces as shown in the experimental results of the paper.

- To demonstrate the generality and reproducibility of the proposed method, we carried out extensive tests in two challenging application domains: i) food image analysis and ii) 3D hand pose estimation. Experimental results with four well-known publicly available datasets and comparison with over fifteen state-of-the-art approaches show the great potential of the proposed method.

The rest of this paper is organized as follows: Section 2 discusses related works in cross-modal approaches, while in Section 3 the proposed framework is presented in detail. Finally, Section 4 presents the experimental set-up and comparisons of the proposed framework against state-of-the-art approaches, while conclusions are drawn in Section 5.

2. Related Work

In this section we present cross-modal frameworks utilizing neural networks as encoders and decoders. The frameworks have been categorized into discriminative and generative. The proposed framework falls under the second category.

Regarding discriminative approaches, Li et al. [11] extract features independently from RGB and depth modalities and an attention mechanism is applied to the fused features, for the task of object detection. For the same task, in [4, 18, 29], the authors leverage semantic knowledge, obtained from textual sources, in order to map images into a rich semantic embedding space. Another approach proposed by Aytar et al. [1] aims to learn cross-modal scene representations for the task of zero-shot recognition and retrieval. Specifically, they regularize a cross-modal CNN to get a joint embedding for different modalities, such as different visual domains and text. The joint representation is initially acquired from a CNN and sentence embeddings are mapped to it. In [17], the authors train a deep autoencoder for cross modality (video and audio) feature learning, to reconstruct both modalities and thus locate correlations across them. Cai et al. [2] employ a depth regularizer during training to improve the RGB-based method, exploiting depth information. Salvador et al. [27] and Carvalho et al. [3] employ a CNN model for obtaining image representations and RNN models for obtaining recipe ingredient and cooking instruction representations, which are utilized for image-to-recipe and recipe-to-image retrieval. More recently, Salvador et al. [26] presented a new framework for recipe generation from food images in which a CNN processes input images and extracts relevant representations. Then ingredient and instruction decoder modules, utilizing the same attention mechanism as in [32], convert these representations into recipe ingredients and cooking instructions.

Concerning generative frameworks, Mueller et al. [14] regress heatmaps for each joint and use a kinematic 3D hand model to predict the 3D hand pose. Liong et al. [33] employ a variational method for cross-modal multimedia retrieval. A fusion network learns to produce binary codes, by processing images and text. Subsequently, they train two variational networks to produce the same code as the fusion network. In this fashion, they achieve to encode both the multi-modal and the single-modality data into the same representation. In [28] the authors use a variational framework, consisting of an encoder-decoder pair for each modality, for the task of zero-shot and few-shot learning. At early stages, they use encoder-decoder pairs for autoencoding purposes, while afterwards they augment the training procedure with alignment between different modalities and distributions. Zhu et al. [39] employ the same network architecture as [27] in order to produce recipe and image embeddings of the same size. Then an unconventional GAN architecture with one generator and two discriminators is used for improving the aforementioned embeddings. The generator is trained to (re)construct images from either recipe or image embeddings. The first discriminator is trained to distinguish between real and generated images, while the second between images generated from image embeddings and recipe embeddings. The full architecture is used for image-to-recipe and recipe-to-image retrieval. Moving in the same direction, Hao et al. [35] also use the base architecture of [27] in order to obtain recipe and image embeddings, which are aligned using a discriminator component to distinguish between them. The proposed architecture also employs these embeddings for retrieval and cross-modal translation. This last objective uses recipe embeddings for generating food images through a GAN as well as image embeddings for recovering ingredients.

As far as VAE-based approaches are concerned, Wan et al. [34] use a combination of GANs and VAEs in order to create two latent spaces, one for depth images and the second one for hand poses. Afterwards they employ a network to map the uniform distribution used in GAN into the normal distribution generated by VAE. Spurr et al. [30] introduce a VAE-based method that leverages different input modalities, such as RGB and depth for hand pose estimation. They encode each modality to a shared latent space and decode a drawn sample to the respective modality. Additionally, they enforce all modality embeddings to lie in the same space, by alternating between decoding into different modalities. Yang et al. [37] follow a similar approach with one unified latent space. They propose a disentangled VAE (dVAE) in order to learn similar latent rep-

resentation that can disentangle poses and other factors, like viewpoint, background, etc. In contrast, the authors in [36] utilize one latent space per input modality and provide two ways of aligning these latent spaces, via KL divergence loss and product of Gaussian Experts to improve the recognition results. In this paper, we propose a novel cross-modal variational alignment of the probability distributions of latent spaces corresponding to different modalities, in order to further improve the overall performance of the network. Compared to similar approaches in the literature, such as [34], where a single neuron is used for aligning two distributions of the same dimensionality, our method employs a VAE network that aligns the distributions through a mapping to an intermediate distribution.

3. Method Description

Given two data modalities M_1 and M_2, our goal is to find an effective approach in order to transition from M_1 to M_2. One such approach consists of traditional encoder/decoder modules, where an input sample from M_1 is encoded into a fixed point in latent space and then decoded into the other modality M_2. Another approach involves their variational counterparts, in which the input from M_1 is encoded into a probability distribution and a sample from this distribution is decoded into M_2. While both of these approaches produce satisfactory results, it would be beneficial to the overall performance of the architecture if extracted information from both modalities M_1 and M_2 could be used for transitioning into M_2, but in such a way as to require only modality M_1 during evaluation.

To this end, the proposed architecture, illustrated in Figure 1, consists of three distinct variational branches in order to accomplish this goal. The upper branch transitions from modality M_1 to M_2, therefore learning a mapping of M_1 into a distribution in a way that is aligned to the final goal. The lower branch is a VAE network that maps M_2 into itself, thus learning to project onto a distribution in a way that favors the reconstruction process. The variational alignment branch, in order to effectively combine the information extracted by the other branches and improve the overall performance, learns to align the distributions produced by E_1 and E_2 in accordance with the target goal, acting as a translation mechanism between the two. An illustration of the effectiveness of the variational alignment branch (mapper), when applied to the domain of food image analysis, can be seen in Figure 2. Food images (M_1) are encoded through both the image encoder E_1 and the mapper VE/VD, while ingredients (M_2) are encoded by the ingredient encoder E_2 and their projections are visualized using t-SNE. The image projections through the mapper network (Os) are much closer to the ingredient projections (Xs) than the image projections produced by E_1 (triangles).

The training process for the architecture takes place in

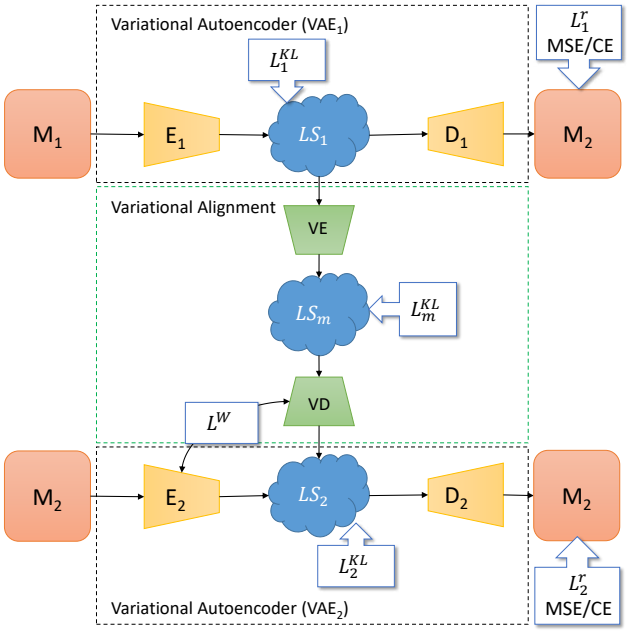

Figure 1. The proposed variational alignment architecture. The upper branch transitions from modality M_1 to M_2 using encoder E_1 and decoder D_1. The lower branch autoencodes M_2 through encoder E_2 and decoder D_2. The middle branch aligns the distribution produced by E_1 to the one produced by E_2 using the variational encoder (*VE*) and decoder (*VD*), which map to and sample from an intermediate distribution.

two distinct phases. The reason for this is that we want the upper and lower variational branches to converge before the variational alignment branch begins the aligning process. During the first phase, the upper and lower branches are trained in parallel, independently of each other. As shown in Figure 1, the encoder E_i maps modality M_i into latent space LS_i and then a sample from this distribution is decoded by D_i into modality M_2, for both networks ($i \in \{1, 2\}$). The encoder E_i produces fixed-size vectors $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ as output, with dimensionality d_i, which parametrize a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i = diag(\sigma_i(1)^2, \ldots, \sigma_i(d_i)^2)$, from which a sample $\boldsymbol{z}$ is drawn. This sample then becomes the input to the decoder D_i which transitions to modality M_2.

The weights of these branches are optimized according to two objectives. The first one is that the produced $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ vectors of E_i match those of a standard normal distribution, by minimizing their Kullback–Leibler divergence [10]:

$$L_i^{KL} = \frac{1}{2} \sum_{j=1}^{d_i} \left(\sigma_i^2(j) + \mu_i^2(j) - \ln \sigma_i^2(j) - 1 \right) \quad (1)$$

The second objective depends on the task at hand. In the case of classification tasks the objective is that the produced

label distribution $\hat{y}$ matches the true label distribution y, by minimizing their cross-entropy [16]:

$$L_i^{CE} = -y \log \hat{y}_i - (1-y) \log (1-\hat{y}_i) \tag{2}$$

In the case of regression tasks the objective is expressed by the MSE loss:

$$L_i^{MSE} = \frac{1}{K} \sum_{j=1}^{K} (y(j) - \hat{y}_i(j))^2 \tag{3}$$

where K is the dimensionality of modality M_2.

After these two branches have finished training, the second training phase begins. During this phase, only the variational alignment branch (mapper) is trained, while both previous branches remain frozen. To this end, E_1 encodes data samples from M_1, producing vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\sigma}_1$. These vectors constitute the input to the mapper, which essentially performs a re-parametrization of the distribution produced by E_1, through a mapping to an intermediate distribution. The distribution parametrized by the mapper-generated $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}_m$ is then used in order to draw a sample $\boldsymbol{z}$, which becomes the input to the decoder of the lower branch D_2. During this stage, in addition to the previous optimization objectives, the variational alignment branch also optimizes the Wasserstein distance [6] between the re-parametrized distribution and the one produced by the encoder E_2:

$$\begin{aligned} L^W = \Big(& \|\boldsymbol{\mu_m} - \boldsymbol{\mu_2}\|^2 + tr(\boldsymbol{\Sigma_m}) + tr(\boldsymbol{\Sigma_2}) \\ & -2tr\big[(\sqrt{\boldsymbol{\Sigma_m}}\boldsymbol{\Sigma_2}\sqrt{\boldsymbol{\Sigma_m}})^{1/2}\big] \Big)^{1/2} \end{aligned} \tag{4}$$

This can be further simplified, due to the covariance matrices being diagonal:

$$L^W = \Big(\|\boldsymbol{\mu_m} - \boldsymbol{\mu_2}\|^2 + \|\boldsymbol{\sigma_m} - \boldsymbol{\sigma_2}\|^2 \Big)^{1/2} \tag{5}$$

The aim of this objective is to better align the distribution produced by the mapper to the one produced by E_2, since decoder D_2 was trained with samples from the latter.

Finally, after the end of the second training phase, a fine-tuning process follows where all network parameters are optimized at the same time. The final loss function that the network optimizes takes the following form:

$$L = \sum_{i=1}^{2} \left(w_i^r L_i^r - w_i^{KL} L_i^{KL} \right) - w_m^{KL} L_m^{KL} + w^W L^W \tag{6}$$

where r can be CE or MSE.

Since training is performed in phases, during the first phase $w_m^{KL} = w^W = 0$, in the second phase where the variational alignment network is trained $w_1^r = w_i^{KL} = 0,\ (i \in \{1, 2\})$, while during fine-tuning all weights have non-zero values.

Due to the generic nature of the proposed method, modalities M_1 and M_2 can be of any type. In this work, two modality pairs (M_1, M_2) have been studied, spanning two different application domains: a) food images and ingredients (presented in Section 4.1) and b) RGB images of hand poses and their corresponding 3D coordinates (presented in Section 4.2).

4. Experimental Evaluation

To evaluate the proposed methodology and at the same time demonstrate the generality of our approach, we conducted extensive tests in two different and very challenging application domains which have attracted recently a lot of interest: i) food image analysis and ii) 3D hand pose estimation. In the case of food image analysis, we focused on the task of ingredient recognition from a single image. More specifically, in this task we considered two modalities, food images and text containing the ingredients, and applied variational alignment in order to map the latent space of food images into that of text. On the other hand, in the second application domain, we focused on the estimation of 3D hand pose configurations from images using a VAE network to initially create the latent space of RGB images and then map it into that of 3D hand pose configurations. Here we have to note that in this task, one could also add more modalities (e.g., depth information), apart from RGB images, in order to further improve the results. However, the alignment of multiple latent spaces into a single latent space is out of the scope of this paper. Experimental results with four well-known datasets and comparison with a number of state-of-the-art approaches show the superiority of the proposed approach.

4.1. Food Image Analysis

4.1.1 Datasets

In order to evaluate the proposed framework for the task of ingredient recognition, the publicly available Yummly-28K [13] and Recipe1M [27] datasets were used. Yummly-28K consists of $27,638$ recipes and images. Custom preprocessing scripts were developed for this dataset so as to ignore preparation methods, measuring units, etc. and only extract food ingredients from the recipe text. In the end, after grouping together similar ingredients (e.g., spaghetti and macaroni) and discarding very rare ones, 265 unique ingredients were identified. This process was necessary since Yummly-28K has not been used in the literature for this task, i.e. ingredient recognition. The dataset was randomly split into 85% for training and 15% for testing, resulting in $23,493$ recipes for training and $4,145$ for testing. This dataset was used for evaluating the performance of the proposed method against other variational frameworks. Recipe1M consists of $252,547$ recipes for training, $54,255$

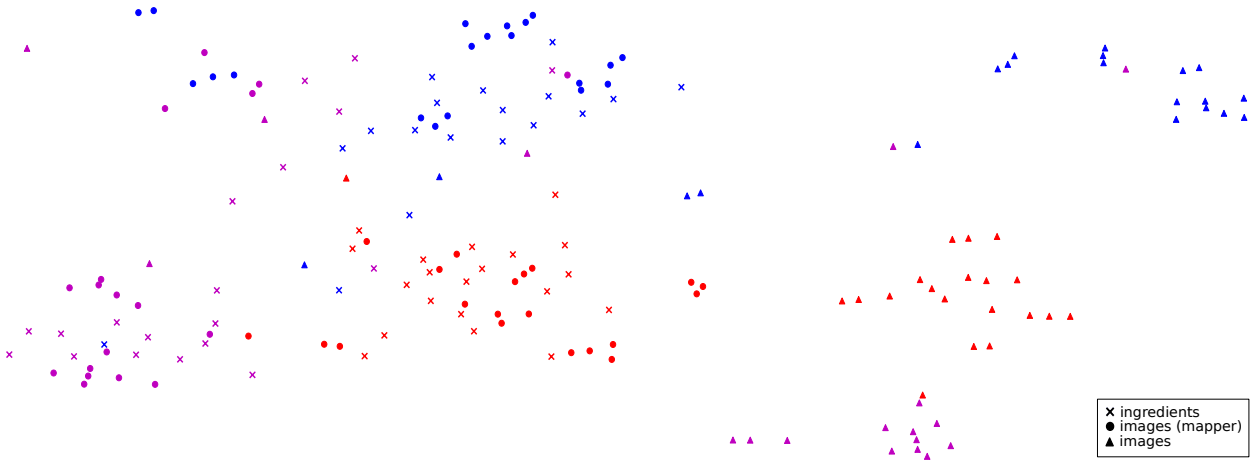

Figure 2. A t-SNE visualization of recipe projections from the image encoder (triangles), ingredient encoder (Xs) and mapper (Os) components. Colours indicate different food categories. It is evident that image projections using the proposed mapper are better aligned to the ingredient ones than image projections without variational alignment.

for validation and $54,506$ for testing, following the preprocessing of [26]. In this dataset there may be more than one image per recipe, while the total number of unique ingredients is $1,488$. In Recipe1M the proposed method was evaluated against current state-of-the-art approaches in ingredient recognition. Two recipe examples from Yummly-28K can be seen in Figure 3.

4.1.2 Implementation Details

In Yummly-28K all images were resized to 360×240, the dimension with the highest frequency of occurrence, since all images have the same aspect ratio. In Recipe1M images were resized to 256 in their shortest side. Following [26], random crops of 224×224 were used during training, while a central crop of 224×224 was used during testing for both datasets. We also adopted the data augmentation scheme of [25], horizontally flipping images with $p = 0.5$ and randomly rotating by ± 10 degrees. During testing no data augmentation was performed, unless indicated by TTA (test-time augmentation) next to the method name.

Regarding the components of the proposed framework, the image encoder E_1 is a DenseNet-121 model with two additional convolutional and average pooling layers before the feed-forward fully-connected layer. This component is pre-trained on ImageNet. The ingredient encoder E_1, both ingredient decoders D_1 and D_2, as well as the variational encoder *VE* and decoder *VD* components, are single-layer fully-connected networks. The dimensionality of all latent probability distributions was set to $d = 512$.

For comparison purposes on Yummly-28K, we have implemented two other cross-modal VAE frameworks: CM-VAE based on [30] and CADA-VAE based on [28]. CM-VAE consists of an image encoder and decoder, as well as an ingredient encoder and decoder, trained similarly to Var. 4 of [30] ($img \rightarrow img$, $ingr \rightarrow ingr$, $img \rightarrow ingr$). The image decoder was implemented as a 7-layer CNN model, while the other components were the same as the ones in our proposed framework. For CADA-VAE the training process involved all four possible paths ($img \rightarrow img$, $ingr \rightarrow ingr$, $img \rightarrow ingr$, $ingr \rightarrow img$). As proposed in [28], the image encoder and decoder processed feature vectors extracted from images instead of the images themselves, and were implemented as single-layer fully-connected networks. Feature vectors were extracted using the same DenseNet-121 architecture mentioned previously. The ingredient encoder and decoder were fully-connected networks as well, the same as our proposed method. An additional classifier was trained in order to provide the final ingredient recognition, as in [28], again implemented as a single-layer FC network.

All methods were trained using the Adam optimizer [9] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 10^{-4}, which was scaled by 0.99 after every epoch. Performance was measured using the F1 score (harmonic mean of precision and recall) and the Intersection over Union (IoU) metrics, applied to the lists of ground truth and predicted ingredients. In Yummly-28K we computed per-recipe F1 and IoU and averaged the results at the end. In Recipe1M they were computed using the code[1] provided by [26].

4.1.3 Experimental Results on Yummly-28K

As can be seen in Table 1, where the ingredient recognition results on Yummly-28K are presented, CM-VAE achieved

[1] https://github.com/facebookresearch/inversecooking

1 whole turkey
2 cups maple syrup
1 cup bourbon
1 tablespoon pickling spice
1 large carrot, scraped
1 celery rib
1 medium onion

chicken breast
seasoned salt
BBQ sauce
cheese slices

Figure 3. Two recipes consisting of images and corresponding ingredients from Yummly-28K.

an F1 score of 39.80 and an IoU of 26.35. CADA-VAE, which employs an explicit distribution alignment objective, was able to improve upon these results by 0.95 in F1 and by 0.8 in IoU. The next two approaches on the table, indicated by + Mapper, refer to augmented versions of the CM-VAE and CADA-VAE architectures with the proposed mapper component. In order to showcase the effectiveness of this component, it was added after the aforementioned architectures had finished training and their weights were frozen. In other words, if we remove the mapper component, i.e., the proposed variational alignment framework, from the CM-VAE + Mapper model, its performance is the same as CM-VAE, since the image encoder and ingredient decoder components were frozen in CM-VAE + Mapper; only the mapper component was able to adjust its weights. We see than in both cases the addition of the mapper improved the results of the baseline methods. CM-VAE + Mapper was 0.77 and 0.65 ahead in terms of F1 and IoU compared to its baseline, while CADA-VAE + Mapper improved the baseline results by 0.25 in F1 and by 0.3 in IoU metrics. The proposed approach was able to outperform CM-VAE by 5.79 points in F1 and 5.26 points in IoU and CADA-VAE by 4.84 F1 and 4.46 IoU points. Employing test-time augmentation provided an increase of 1.64 and 1.36 in F1 and IoU respectively, compared to the baseline.

Method	F1	IoU
CM-VAE	39.80	26.35
CADA-VAE	40.75	27.15
CM-VAE + Mapper	40.57	27.00
CADA-VAE + Mapper	41.00	27.45
Proposed	**45.59**	**31.61**
Proposed + TTA	**47.23**	**32.97**

Table 1. Ingredient recognition results of various cross-modal variational methods on Yummly-28K.

4.1.4 Experimental Results on Recipe1M

On the large-scale Recipe1M dataset, the proposed method was evaluated against state-of-the-art approaches in ingredient recognition from food images. The experimental results are shown in Table 2. The first two approaches, R_{I2L} and R_{I2LR} [27], are retrieval-based, so their predictions correspond to the ingredients of the closest-matching recipe. Reported results for these methods are from [26]. The next two models are non-variatonal and have feed-forward fully-connected (FF_{TD}) and transformer-based (TF_{set}) classifiers [26]. As we can see, the proposed method outperformed the retrieval approaches by a large margin, namely by 17.35 and 16.05 points in terms of F1. The proposed method surpassed the FF_{TD} approach, which has the same classifier, by 3.24 F1 and 2.79 IoU points, while it managed to outperform even the TF_{set} approach with the transformer-based classifier by 0.57 and 0.5 in terms of F1 and IoU. Test-time augmentation provided further improvements to the recognition rate of the proposed framework, widening the difference to the transformer-based network to 1.44 F1 and 1.27 IoU points.

Method	F1	IoU
R_{I2L} [27]	31.83	18.92
R_{I2LR} [27]	33.13	19.85
FF_{TD} [26]	45.94	29.82
TF_{set} [26]	48.61	32.11
Proposed	**49.18**	**32.61**
Proposed + TTA	**50.05**	**33.38**

Table 2. Ingredient recognition results on Recipe1M.

4.2. 3D Hand Pose Estimation

4.2.1 Datasets

In regard to the task of 3D hand pose estimation, our method is evaluated on two publicly available datasets, Rendered Handpose Dataset (RHD) [40] and Stereo Hand Pose Tracking Benchmark (STB) [38]. RHD is a synthetic dataset containing rendered hand images from 20 characters performing 39 actions. It consists of 41258 images for training and 2728 images for evaluation, with 320×320 resolution. For each sample both 3D and 2D hand pose, as well as depth map and segmentation mask are provided. This dataset is higlhy challenging as it contains heavily occluded fingers, visual diversity and noise. STB includes 12 sequences with 6 different backgrounds of finger counting and random poses. Each of these sequences consists of 1500 frames,

with a resolution of 640×480, resulting in 18k samples, 15k for training and 3k for testing. 3D keypoint annotations are provided and consequently a camera intrinsic matrix can be utilized to obtain 2D keypoint locations.

In order to evaluate the performance of our proposed method, we use the two most common metrics on hand pose estimation field, mean End-Point-Error (EPE) and Area Under the Curve (AUC) on the Percentage of Correct Keypoints(PCK). Mean EPE measures the average euclidean distance between ground-truth and predicted keypoints, while PCK is the mean percentage of predicted keypoints below different error thresholds, in comparison to the correct keypoint location.

4.2.2 Implementation Details

We performed the same data pre-processing as previous works [40, 30, 37, 8], to be directly comparable. We utilized 2D annotations in both datasets to create a bounding box around the hand region. Afterwards, we randomly rotated it in the range $[-45°, 45°]$, applied random vertical flip with $p = 0.5$ and resized the image to 256×256. At test time, no data augmentation was conducted and the bounding box was resized to 256×256. In addition handedness, palm center and scale of the hand were provided during both training and testing. Since RHD provides the location of the wrist-joint, while in contrast STB gives the location of the palm-joint, we shifted the wrist-joint in RHD into the palm one, in order to make annotations on both datasets consistent, following [40, 30]. Importantly, as [12] indicated, self-occlusion of the hand results in different observations for the same pose. For that reason, we moved the center of the hand to the center of the bounding box and accordingly rotated the 3D pose. Thus the 3D centroid of the hand aligned with the camera's z-axis and the one-to-many mapping for the image-pose pairs was alleviated. This is the same procedure as in [36].

Resnet-18 [7] was employed as E_1 to encode RGB images. We adjusted the last fully connected layer such as to predict the mean and variance of a normal distribution for a given sample. As far as 3D hand pose encoder, E_1, and decoders, D_1 and D_2 are concerned, we used 6 fully connected layers, with 512 units per layer, while each of the mapper components consists of a single fully connected layer. In our experiments we set batch size to 64 for RHD and to 32 for STB, as it is considerably smaller, and used the Adam optimizer with a learning rate of 10^{-4}, which was scaled by 0.99 after each epoch.

4.2.3 Experimental Results on RHD and STB

In this section, we compare the proposed method against a number of hand pose estimation state-of-the-art methods. Initially, we use the mean EPE metric to compare our method with other RGB-to-3D methods, i.e. Zimmerman et al. [40], Spurr et al. [30] and Yang et al. [37] using both RHD and STB datasets. The experimental results of this comparison are presented in Table 3. As we can see, the proposed method outperforms all other methods providing 15.61 and 6.93 mean EPE on RHD and STB datasets, respectively. More specifically, the proposed cross-modal variational alignment approach achieves improvements up to 4.12 and 4.34 mean EPE in RHD and 1.63 and 1.73 mean EPE in STB from the cross-modal deep Variational Autoencoder [30] and the disentangled Variational Autoencoder (dVAE) [37] respectively, showing its great potential to align the two latent spaces and boost significantly the network's performance. As we mentioned in the introduction of this section, there are other works, e.g., [36, 5, 8], that utilize multiple modalities during training, such as cloud points, heatmaps and depth maps, guiding their network to benefit from different latent spaces, however, the alignment of multiple latent spaces into a single latent space is out of the scope of this paper and can be studied in a future work.

Method	RHD	STB
Zimmerman et al. [40]	30.42	8.68
Yang/Yao et al. [37] *	19.95	8.66
Spurr et al. [30] *	19.73	8.56
Proposed *	**15.61**	**6.93**
** VAE-based methods*		

Table 3. 3D hand pose recognition results.

Finally, we compare the PCK curves with a number of state-of-the-art methods on both datasets. More specifically, as shown in Figure 4, our framework surpasses all state-of-the-art methods [40, 30, 37] on RHD dataset performing 0.907 AUC. Similarly, as illustrated in Figure 5, the proposed method achieves the highest AUC score, i.e., 0.997, on STB dataset outperforming all other existing methods [40, 21, 30, 14, 37, 2, 19, 23, 31]. Figure 6 visualizes several pose predictions of our method compared to plain RGB-to-3D network, in order to demonstrate the crucial contribution of our cross-modal variational alignment approach. Results prove that our framework is capable of mitigating the challenges of this task, such as self-occlusions, and consequently improves recognition performance.

5. Conclusions

In this work a novel approach for cross-modal variational alignment of latent spaces was presented. The proposed approach aligns the latent distributions of two VAE models, processing modalities M_1 and M_2, through the use of an additional VAE component that explicitly incorporates the minimization of their distance into its optimization objectives. The generic nature of this approach allows it to be

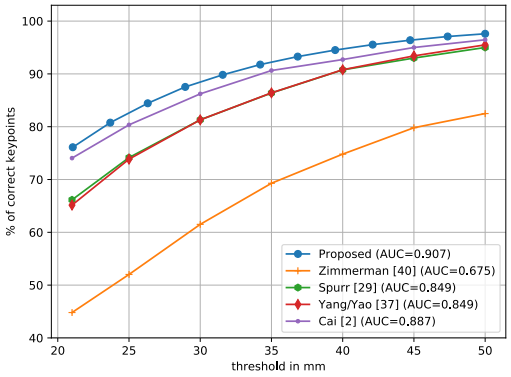

Figure 4. AUC on PCK curve: Comparison to state-of-the-art methods on RHD

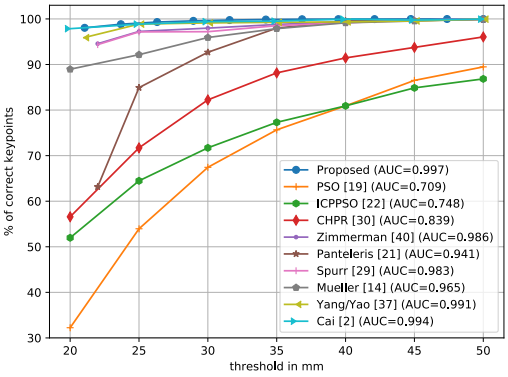

Figure 5. AUC on PCK curve: Comparison to state-of-the-art methods on STB

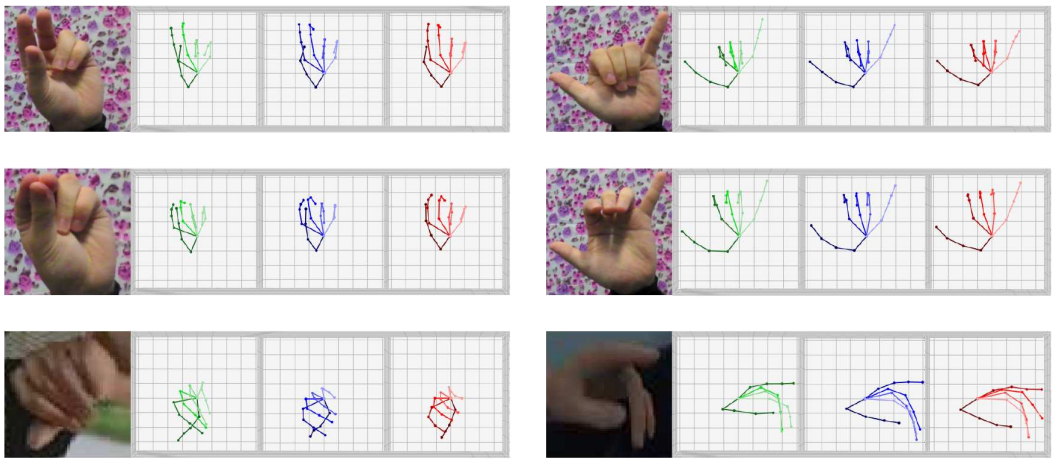

Figure 6. A comparison between 3D poses. Left column depicts ground-truth joints (green), our framework's predictions are shown in middle column (blue) and poses acquired without the mapper branch are in right column (red).

applied to any type of data modality pairs. Experimental results in two different and very challenging application domains have demonstrated the effectiveness of our method compared to state-of-the-art approaches. As future work, the proposed framework could be extended to provide aligned representations for more than two modalities.

Acknowledgment. This work has been supported from EC under grant agreement no. H2020-ICT-19-2016-2 "EasyTV: Easing the access of Europeans with disabilities to converging media and content".

References

[1] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2303–2314, 2017.

[2] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *European Conference on Computer Vision*, pages 666–682, 2018.

[3] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *Research & Development in Information Retrieval*, pages 35–44, 2018.

[4] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[5] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Computer Vision and Pattern Recognition*, pages 10833–10842, 2019.

[6] Clark R Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *European Conference on Computer Vision*, pages 118–134, 2018.

[9] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *CoRR abs/1412.6980*, 2014.

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[11] Guanbin Li, Yukang Gan, Hejun Wu, Nong Xiao, and Liang Lin. Cross-modal attentional context learning for rgb-d object detection. *IEEE Transactions on Image Processing*, 28(4):1591–1601, 2018.

[12] Shile Li and Dongheui Lee. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In *Computer Vision and Pattern Recognition*, pages 11927–11936, 2019.

[13] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia*, 19(5):1100–1113, 2016.

[14] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Computer Vision and Pattern Recognition*, pages 49–59, 2018.

[15] Meinard Müller. Music synchronization. In *Fundamentals of Music Processing*, pages 115–166. Springer, 2015.

[16] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[17] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning*, pages 689–696, 2011.

[18] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[19] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *British Machine Vision Conference*, volume 1, page 3, 2011.

[20] Charles B Owen and Fillia Makedon. Cross-modal retrieval of scripted speech audio. In *Multimedia Computing and Networking 1998*, volume 3310, pages 226–235, 1997.

[21] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *Winter Conference on Applications of Computer Vision*, pages 436–445. IEEE, 2018.

[22] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019.

[23] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Computer Vision and Pattern Recognition*, pages 1106–1113, 2014.

[24] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *International Conference on Multimedia*, pages 251–260, 2010.

[25] Amaia Salvador. *Computer Vision beyond the visible: Image understanding through language*. PhD thesis, UNIVERSITAT POLITÈCNICA DE CATALUNYA, 2019.

[26] Amaia Salvador, Michal Drozdzal, Xavier Giro-i Nieto, and Adriana Romero. Inverse cooking: Recipe generation from food images. In *Computer Vision and Pattern Recognition*, pages 10453–10462, 2019.

[27] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Computer Vision and Pattern Recognition*, pages 3020–3028, 2017.

[28] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *Computer Vision and Pattern Recognition*, June 2019.

[29] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013.

[30] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Computer Vision and Pattern Recognition*, pages 89–98, 2018.

[31] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *Computer Vision and Pattern Recognition*, pages 824–832, 2015.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[33] Liong Venice Erin, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Cross-modal deep variational hashing. In *International Conference on Computer Vision*, pages 4077–4085, 2017.

[34] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *Computer Vision and Pattern Recognition*, pages 680–689, 2017.

[35] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Computer Vision and Pattern Recognition*, pages 11572–11581, 2019.

[36] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *International Conference on Computer Vision*, pages 2335–2343, 2019.

[37] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *Computer Vision and Pattern Recognition*, pages 9877–9886, 2019.

[38] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016.

[39] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Computer Vision and Pattern Recognition*, pages 11477–11486, 2019.

[40] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *International Conference on Computer Vision*, pages 4903–4911, 2017.