APPENDIX: Complementary-Contradictory Feature Regularization against Multimodal Overfitting

Antonio Tejero-de-Pablos
CyberAgent
Shibuya, Tokyo, Japan
antonio_tejero@cyberagent.co.jp

1. Grounding on the Digits dataset

As in previous multimodal learning works, we grounded the bimodal setup via the *Digits* dataset.

Digits [15] combines datasets of images that represent the same digit in different styles (*i.e.*, MNIST and SVHN), for the task of digits classification. Input images (size 32×32) on both modalities are paired and assigned a digits label from 0 to 9 ($J = 10$). VQVAE contains 256 codes of size $D = 32$. Training is run for 10 epoch on the train-test splits defined in [7].

Table 1 shows the quantitative evaluation on the *Digits* dataset. A large ratio of the complementary features is made up of MNIST ($m = 1$) codes, while most of the SVHN ($m = 2$) codes remain contradictory. This is reasonable since MNIST contains all information required to classify a digit, whereas SVHN images contain lots of unrelated information such as background and colors. Thus, their ζ value is very low, as one modality does not add much useful information to the other. These results show the difference between previous shared/private spaces and our complementary/contradictory disentanglement: while multimodal generation tasks leverage redundant information (*e.g.*, the color red and the word "red") to generate their shared space, in classification tasks, redundant features are not considered complementary.

As for the qualitative evaluation, Fig. 1 (a) shows the reconstructions of the input modalities, for all the features $\psi(\hat{z}_i^m|\theta) = \tilde{x}_i^m$, the complementary features $\psi(\hat{z}_i^m \odot \omega^m|\theta)$, and the contradictory features $\psi(\hat{z}_i^m \odot \overline{\omega^m}|\theta)$. We used a colored version of MNIST for a clearer visualization. The complementary space in MNIST shows lines representing the numbers, but features such as line thickness or color variations are not contained, as they are not helpful to solve the task. SVHN modality (b) does not contain complementary features, as the MNIST modality is enough to solve the task. The missing complementary features can be found on the reconstructed contradictory features.

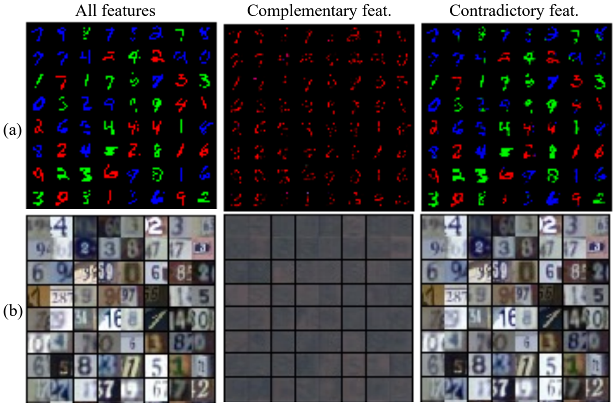


Figure 1. All features, complementary features and contradictory features reconstructed with the learned decoder for modalities (a) MNIST and (b) SVHN in Digits.

This additional experiments aim to show the conceptual differences between our methodology and that of the aforementioned previous works. Methods based on Shared/Private latent spaces [6, 7, 10] separate MNIST and SVHN's content and style features in the shared and private space respectively, while ours only retains MNIST's content features as complementary and leaves style and the whole SVHN modality as contradictory. Also, in [4], it is unclear if interferring color features are still present or not after applying their method, while this can be clearly confirmed in ours.

2. Dataset details

CREMA-D [14] is an image-audio dataset for emotion recognition. It contains 7442 short videos of 91 actors of different nationalities reciting a sentence with emotional content, for the task of emotion recognition ($J = 6$): *angry*, *happy*, *sad*, *neutral*, *disgust*, *fear*. For the visual, we randomly sample a frame from the video and crop the 128×128 region where the face is located. The audio signal is trans-

Table 1. Task accuracy, latent disentanglement and modal complementarity for the Digits dataset.

Dataset	Recons. err. 1	Recons. err. 2	Accuracy	contr^1	contr^2	compl^1	compl^2	ζ
Digits	0.01 (MNIST)	0.01 (SVHN)	99.24%	26.56%	48.44%	23.44%	1.56%	0.06

formed to a 256×256 Mel-Spectrogram using librosa [8], as in [9]. The train-validation-test splits are set as in [9]. Training is run for 100 epochs and the model with the best classification accuracy on the validation set is saved for evaluation.

PennAction [17] is an image-pose dataset for pose recognition. It contains videos for different human actions ($J = 15$) and the positions in image coordinates of 13 body joints, for the task of action recognition. Similar to CREMA-D, we sample and rescale video frames to 256×256. Pose sequences are subsampled/zero-padded to a common length of 100 frames. The train/test splits are those defined in the dataset and used in [1], and training is run for 200 epochs.

NYUv2 [11] is an color-depth dataset for semantic segmentation. It contains 795 pairs of color and depth images for training and 655 for testing. We rescale the color, depth and semantic label images to 256×256. Training is run for 200 epoch with a learning rate of 0.0001 and cosine-decay as in [20]. We solve semantic segmentation as a pixel-wise classification task for the label image, so we keep the same task loss, but we replace the classifier of the task-solver by the decoder architecture used in the VQVAEs (changing the number of output classes).

RML [14] is a video dataset for emotion recognition. It contains 720 samples from 8 actors of different nationalities reciting a sentence with six emotion labels: *angry*, *disgust*, *fear*, *happy*, *sad*, *surprise*.

3. Architectural details of CM-VQVAE

The main text contains the general details of our proposed method, CM-VQVAE, which consists on:

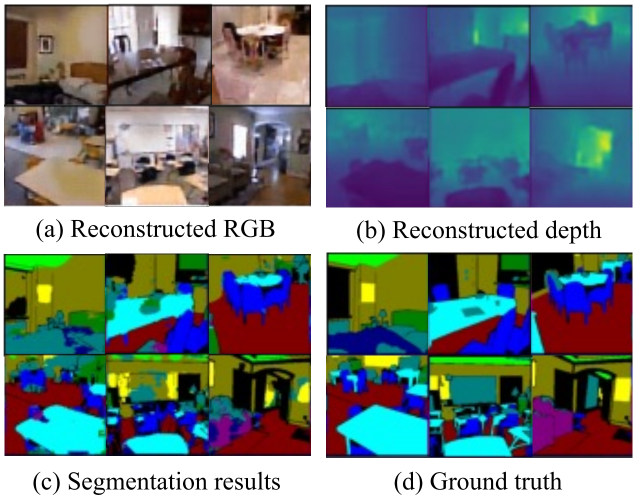
- Modality-specific VQVAE modules, with an encoder φ, a decoder ψ, and a codebook C.
- A Task-solver module, with a mask ω and a classifier γ.

As in the comparison method [9], the architecture of φ, ψ and γ is based on ResNet18, and VQVAEs were adapted from the public implementation of VQVAE[1]. Tab. 2 details the hyperparameters of the convolutional and residual layers, as well as the codebook for each dataset.

We run our model in a GPU NVIDIA A100-SXM (CUDA Version: 11.4). Further architectural details can be found in our code[2].

[1] https://github.com/zalandoresearch/pytorch-vq-vae

[2] https://github.com/CyberAgentAILab/CM-VQVAE


(a) Reconstructed RGB

(b) Reconstructed depth

(c) Segmentation results

(d) Ground truth

Figure 2. Multimodal semantic segmentation results: (a) Color recons., (b) Depth recons., (c) Segmentations and (d) Ground truth in NYUv2.

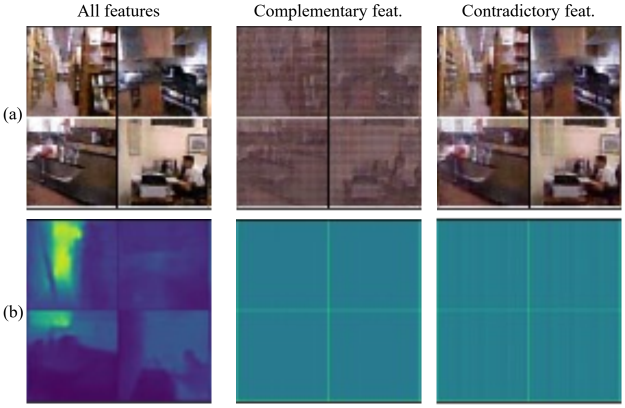


Figure 3. All features, complementary features and contradictory features reconstructed with the learned decoder for modalities (a) color and (b) depth in NYUv2.

4. Qualitative evaluation of NYUv2

Figure 2 shows some visual results of the semantic segmentation task. Fig. 3 displays the reconstructions of the feature spaces in the NYUv2 dataset, obtained as in Fig. 1. As explained in the Discussion (Sec. 5 in the main text), the features disentangled in the latent spaces does not necessarily have to have a semantic meaning, and this dataset is an example.

Table 2. Architectural details of CM-VQVAE for each dataset.

Dataset	**#channels convolutions**	**#channels residuals**	**#layers residuals**	**Size of z_i^m** $(H \times W \times D)$	**#codes VQVAE** K	**VQVAE commit. cost**	**Batch size**
CREMA-D	128	32	2	$32 \times 32 \times 64$	512	0.25	128
PennAction	128	32	2	$32 \times 32 \times 64$	512	0.25	128
NYUv2	128	32	2	$32 \times 32 \times 64$	512	0.25	8
Digits	64	12	1	$16 \times 16 \times 32$	256	0.25	128

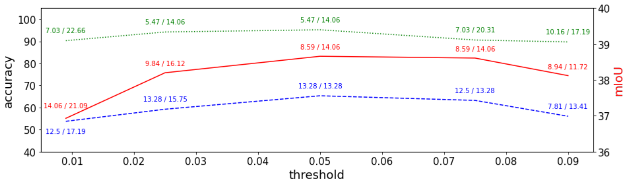


Figure 4. Accuracy (left y axis) and mIoU (right y axis) vs. t in CREMA-D (dashed blue), PennAction (dotted green) and NYUv2 (red) datasets. Labels indicate compl^{color} / compl^{audio}, compl^{color} / compl^{pose} and compl^{color} / compl^{depth}.

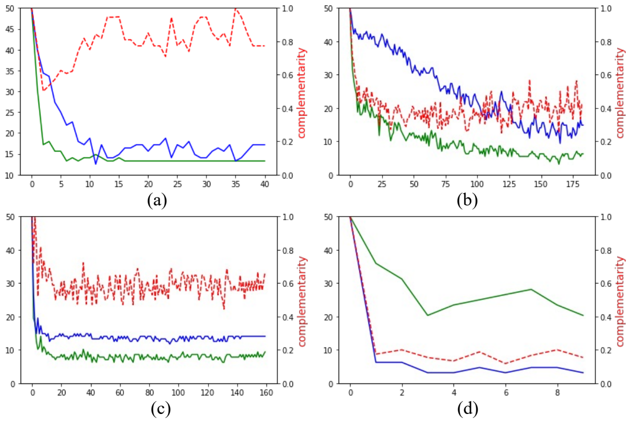
5. Comparison with SOTA architectures

Table 3 compares our method with other state-of-the-art architectures for each task. Applying our proposed multimodal learning method to a generic ResNet architecture outperforms most approaches for all tasks. This comparison is just a reference, as some methods use specific architectures and learning curricula (*e.g.*, Transformers, pretraining, etc.), and similarly, our accuracy could be further improved by changing the ResNet backbone.

6. Threshold t

The value of t was chosen empirically. Fig. 4 shows the effect on the task performance when varying t. High ts entail masking too many features from the start, which hinders learning and interaction among modalities. On the other hand, low ts result in not masking enough features. Both cases converge consistently to a suboptimal disentanglement in our experiments. Our method is not highly sensitive to t values around the optimal range, but there is a significant performance gap for extreme values of t. The reason is that no masking or forcibly masking features is equivalent to not applying our method (see ablation on Tab. 3 in the main text). Note that, unlike other hyperparameters in our method, t controls learning as *e.g.* the learning rate does, and thus, unreasonable values can hinder the task performance.

In addition, the same value t is used for all modalities, since setting different ts would prevent multimodal interaction during feature selection. The threshold t represents the criterion to mask multimodal features altogether simultaneously. This allows selecting *e.g.* certain audio features over color features and viceversa. Moreover, different ts would prevent ζ from measuring the contribution of one modality over the other, as the criterion for selecting features would be different. Instead, the mask weights are learnable so, under the same criterion, the network decides if the weight value for a certain modality feature should be over or below t.



(a)

(b)

(c)

(d)

Figure 5. Evolution during training epochs (x-axis) of the sizes compl^m (left y-axis) and complementarity ζ (right y-axis). In CREMA-D (a), (green) is color and (blue) is audio. In PennAction (b), (green) is color and (blue) is pose. In NYUv2 (c), (green) is color and (blue) is depth. In Digits (d), (green) is MNIST and (blue) is SVHN. In all of them, (dotted-red) is ζ.

7. Complementarity measure

7.1. Complementarity during training

We monitor how complementarity evolves during training (Fig. 5). Initially, both modalities have compl$^m = 50\%$ (*i.e.*, contr$^m = 0\%$). In CREMA-D (a), both modalities present a high complementarity, which increases as color and audio features are optimized. In PennAction (b), color features are rapidly discarded, which results in a low complementarity since the beginning. In NYUv2 (c), complementarity increases very slowly due to the learning schema of the dataset (*i.e.*, low learning rate with cosine decay scheduling in order to learn the more imbalanced classes). In Digits (d), the first epochs relegate most of the irrele-

Table 3. Comparison with the related work.

CREMA-D	**Acc.**(%)
Grad-Blend [13]	56.8
OGM-GE [9]	57.7
PMR [4]	61.1
Ours	**65.32**

PennAction	**Acc.**(%)
HDM-BG [18]	93.4
C3D [2]	94.3
Transformer [1]	**98.7**
Ours	95.22

NYUv2	**mIoU**
HRNet-18 [12]	33.18
U-Net++ [19]	34.74
MaskSup [20]	38.54
Ours	**38.66**

vant features to the private space, including almost the entire SVHN modality.

7.2. Complementarity in the related work

Some recent works calculate metrics that characterize the processes involved in multimodal learning, in an attempt to add further explainability. Here we provide a brief survey.

In [9], the interaction between two modalities during training is studied using a discrepancy ratio between their predictions on the classification task. In [3], to account for the complementary information that different modalities contribute to a contrastive learning task, inter-modality scores are learned to weight modality-specific features. Modal "relatedness" was introduced in [6], as a measure of how close are the latent representations of the same semantic concept for different modalities (e.g., an image and a caption of a *bird*). This concept is more applicable to generation tasks, but not necessarily to classification, where different modalities usually represent different semantics. In [5], informativeness is defined as the usefulness of individual features and entire modalities to solve a given multimodal task. It is calculated as learnable weights for each feature on a medical table-data database. While defining a "feature unit" in table data (*i.e*., cell values) is straightforward, in multimedia data (*i.e*., image pixels) is not trivial. In [16], complementarity was defined as the information that one modality supplements to other, and leveraged for multimodal domain adaptation. Cross-modal information is modeled between pairs of modalities via a gating operation that weights the features of each modality. As we showed in our experimental results, masking is more effective than simply weighting. In [4], a metric to evaluate the degree of imbalance between two-modalities in real-time is proposed. They calculate the ratio between the Euclidean distance of the each modality's features within a batch and their respective class prototype. Although this gives an idea of the regularization imbalance between multimodal features, it still does not allow quantifying the ratio of useful/irrelevant features of each modality.

Note that each indicator is suitable for a different learning scenario, and comparing them side by side is not reasonable.

8. Extension of our method to $M > 2$

A possible approach would be calculating a ζ^m for each modality as: $\mathrm{compl}^m / \sum_{n=1,n\neq m}^{M} \mathrm{compl}^n$. Also, it would be necessary to determine if single complementarity would be enough (*e.g*., visual-audio-text), or if additional pairwise complementarities should be calculated (*e.g*., visual-audio, visual-text, audio-text).

References

[1] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. STAR-Transformer: A spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 3330–3339, 2023. 2, 4

[2] Congqi Cao, Yifan Zhang, Chunjie Zhang, and Hanqing Lu. Body joint guided 3-d deep convolutional descriptors for action recognition. *IEEE Transactions on Cybernetics*, 48(3):1095–1108, 2017. 4

[3] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5Product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21252–21262, 2022. 4

[4] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2023. 1, 4

[5] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20707–20717, 2022. 4

[6] Tom Joy, Yuge Shi, Philip HS Torr, Tom Rainforth, Sebastian M Schmon, and N Siddharth. Learning multimodal VAEs through mutual supervision. In *Proceedings of the International Conference on Learning Representations*, pages 1–11, 2022. 1, 4

[7] Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal VAE for learning of latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2021. 1

[8] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings*

of the 14th python in science conference, volume 8, pages 18–25, 2015. 2

[9] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 2, 4

[10] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H.S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[11] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*, pages 746–760, 2012. 2

[12] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 4

[13] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 4

[14] Yongjin Wang and Ling Guan. Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia*, 10(5):936–946, 2008. 1, 2

[15] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3964–3973, 2018. 1

[16] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14722–14732, 2022. 4

[17] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2248–2255, 2013. 2

[18] Rui Zhao, Wanru Xu, Hui Su, and Qiang Ji. Bayesian hierarchical dynamic model for human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7733–7742, 2019. 4

[19] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019. 4

[20] Hasib Zunair and A Ben Hamza. Masked supervised learning for semantic segmentation. In *The 33rd British Machine Vision Conference*, 2022. 2, 4