A UNIFIED CONDITIONAL DISENTANGLEMENT FRAMEWORK FOR MULTIMODAL BRAIN MR IMAGE TRANSLATION

Xiaofeng Liu, Fangxu Xing, Georges El Fakhri, Jonghye Woo

Dept. of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

ABSTRACT

Multimodal MRI provides complementary and clinically relevant information to probe tissue condition and to characterize various diseases. However, it is often difficult to acquire sufficiently many modalities from the same subject due to limitations in study plans, while quantitative analysis is still demanded. In this work, we propose a unified conditional disentanglement framework to synthesize any arbitrary modality from an input modality. Our framework hinges on a cycle-constrained conditional adversarial training approach, where it can extract a modality-invariant anatomical feature with a modality-agnostic encoder and generate a target modality with a conditioned decoder. We validate our framework on four MRI modalities, including T1-weighted, T1 contrast enhanced, T2-weighted, and FLAIR MRI, from the BraTS'18 database, showing superior performance on synthesis quality over the comparison methods. In addition, we report results from experiments on a tumor segmentation task carried out with synthesized data.

***Index Terms*—** Image synthesis, Generative Adversarial Networks, Deep learning, Brain tumor

1. INTRODUCTION

Multimodal magnetic resonance (MR) images are often required to provide complementary information for clinical diagnosis and scientific studies [1, 2]. For example, multimodal MR imaging (MRI) with T1-weighted, T1ce (contrast enhanced), T2-weighted, and FLAIR (FLuid-Attenuated Inversion Recovery) MRI can offer greater sensitivity to tumor heterogeneity and growth pattern than single modality, T1ce MRI, thereby benefiting diagnosis, staging, and monitoring of brain metastasis [3]. However, in practice, it is often difficult to acquire sufficiently many modalities due to limitations in study plans, and some modalities could be missing due to imaging artifacts [4, 5].

In recent years, cross-modality synthesis of brain MR images using generative adversarial networks (GANs) has gained its popularity [1]. For example, Yu et al. [5] adopted a pair-wise image-to-image network via Pix2Pix [6, 2] for transferring T1-weighted to either T2-weighted or FLAIR MRI. Also, a cycle-reconstruction approach via CycleGAN for unpaired image translation [7] was introduced in [4, 8] to stabilize the training. These methods [1, 5, 4, 8, 2] aimed at modeling the mapping between two specific modalities, requiring two inverse autoencoders to achieve the cycle-reconstruction [7].

However, the aforementioned approaches have a limitation in that they dealt with the cross-modality synthesis problem which cannot be easily scalable to multiple modalities (i.e., more than two modalities). In other words, in order to learn all mappings among M modalities, $M(M-1)$ different generators have to be trained and deployed (e.g., 12 possible cross-modality networks for T1-weighted, T1ce, T2-weighted, and FLAIR MRI) [9]. Moreover, each translator cannot fully use the entire training data, but can only learn from a subset of data (two out of M modalities). Failure to fully use the whole training data is likely to limit the quality of generated images. To address this, recently, Xin et al. [10] proposed to construct a 1-to-3 network to translate T2-weighted to T1-weighted/T1ce/FLAIR MRI based on Pix2Pix [6]. The improvement over Pix2Pix was achieved by utilizing $3\times$ training pairs for one translator [10]. Besides, the closely related multiple tasks mutually reinforced each other [11]. Yet, with the 1-to-3 network, the number of models to be trained was still limited to the number of modalities.

In this paper, we propose to achieve all of the pair-wise translation using a single set of conditional autoencoder and discriminator. Our framework is scalable to many modalities and can effectively use all of possible paired cross-modality training data. Several unpaired multi-domain synthesis methods are inherently multimodal translation, while they usually require multiple domain-specific autoencoders [12] and discriminators [13], and do not consider pair-wise training data [12, 13, 14, 11]. Without the pair-wise supervision, the largely unconstrained image generation tends to alter important characteristics of an input modality for generating diverse outputs. Unlike image-to-image translation in computer vision, in medical domain, it is of paramount importance not to introduce unnecessary changes and artifacts to carry out quantitative analyses [15].

In addition, these methods ignore the inherent connection between different MR modalities [16, 16]. Since multiple MR modalities are acquired with different scan parameters for the same subject, there should be a shared modality-invariant anatomical feature space [16]. Accordingly, we propose to

configure a pair-wise disentanglement approach [17, 18] to extract an anatomical feature with a modality-agnostic encoder, and then inject a modality-specific appearance with a conditional decoder.

Specifically, our unified multimodal translation framework hinges on the autoencoder, where the decoder is conditioned on the target modality to utilize the paired training data. After the feedforward processing of the autoencoder conditioned on any modality label, the same autoencoder is called again, which conditioned on the modality label of the original input for the cycle-reconstruction. The anatomical information disentanglement is simply enforced by the similarity of the output feature map of the encoder [18, 17].

We empirically validate its effectiveness on the BraTS'18 database, showing superior performance over the comparison methods.

2. METHODOLOGY

Given a set of co-registered M MR modalities, a sample x with modality m_x has $M-1$ pixel-wise aligned samples with the other modalities. The target modality of image synthesis and the corresponding ground truth sample are denoted as m_y and x_y, respectively. Given the pair of input sample and target modality $\{x, m_y\}$, we propose to learn a parameterized mapping $f : \{x, m_y\} \rightarrow \tilde{x}_y$ from $\{x, m_y\}$ to the generated corresponding sample with modality m_y to closely resemble x_y. m_y denotes a four-dimensional one-hot vector to represent the four MR modalities available in the BraTS'18 database. Of note, $m_y = m_x$ indicates the self-reconstruction. The proposed framework is shown in Fig. 1.

2.1. Disentangled United Multimodal Translation

A straightforward baseline network structure for paired two-modality translation is an autoencoder, which is constructed with an encoder Enc and a decoder Dec. Briefly, it first maps x to a latent feature z via the Enc, and then decode z to reconstruct the target image via the Dec. The target ground truth image serves as a strong supervision signal, while the unpaired translation cannot benefit from such regularization [7, 19, 12, 11, 13].

However, the autoencoder has a limitation in that the generated images are likely to be blurry [2], which is partly caused by the element-wise criteria such as the $\mathcal{L}_1$ or $\mathcal{L}_2$ loss [20]. Although recent studies [21] have substantially improved the predicted log-likelihood in the autoencoder, the image generation quality of the autoencoder is still inferior to GANs. In addition, enforcing pixel-wise similarity is likely to distract the autoencoder from understanding the underlying anatomical structure, when inputting slightly misregistered data.

In order to enforce high-level semantic similarity and improve quality of generated textures, recent cross-modality translation models [4, 5, 10] adopted an additional adversarial loss $\mathcal{L}_{adv}$ with a discriminator Dis following Pix2Pix [6, 2],

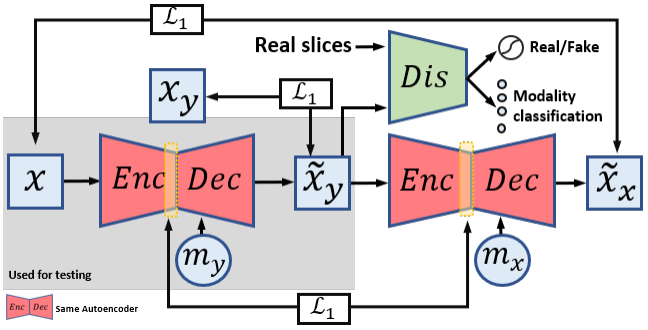

Fig. 1. Illustration of the proposed unified conditional adversarial framework for multimodal co-registered brain MR image translation. Note that only one Enc-Dec set is recalled twice, and only the gray masked subnets are used for testing.

where the training objectives consist of both the $\mathcal{L}_1$ loss and adversarial GAN loss:

$$\min_{Enc,Dec}\max_{Dis} \mathcal{L}_1(\tilde{x}_y, x_y) = |\tilde{x}_y - x_y|, \quad (1)$$

$$\min_{Enc,Dec}\max_{Dis} \mathcal{L}_{adv} = \mathbb{E}\log(Dis(\tilde{x}_y)) + \mathbb{E}\log(1 - Dis(\tilde{x}_y)). \quad (2)$$

To extend the Pix2Pix basenet to multimodal translation, we adopt the conditional decoder structure that takes both the feature map extracted by the encoder $Enc(x)$ and the target modality code m_y as input. The target modality code is spatially replicated and concatenated with the input image. Of note, the unpaired multi-domain synthesis network takes the modality code as input to the encoder [12, 13, 14, 11]; therefore it cannot achieve the disentanglement of modality-agnostic anatomic and modality-specific factors [16, 17, 18]. Then, a single autoencoder model can be switched to all possible pair-wise cross-modality translations. Therefore, $\tilde{x}_y$ in Eqs. (1-2) can be $Dec(Enc(x), m_y)$.

Instead of configuring M Dis for all modalities, we introduce an auxiliary modality classifier Dis_{mc} [22] on top of Dis that allows a single Dis to control multiple modalities. The to be minimized modality classification loss can be formulated as:

$$\mathcal{L}_{mc} = \mathbb{E}_{\tilde{x}_y}[-\log Dis_{mc}(\tilde{x}_y, m_y)] + \mathbb{E}_{x}[-\log Dis_{mc}(x, m_x)]. \quad (3)$$

The objective of conditional GAN with multi-task discriminator can induce an output distribution of over $(\tilde{x}_y|m_y)$ that matches the empirical distribution of real images with modality m_y, i.e., x_y. However, the mapping between two distributions can be largely unconstrained and have many possible translations f to induce the same distribution over $f(\{x, m_y\})$ [7]. Therefore, the learned f cannot guarantee that the individual inputs $\{x, m_y\}$ and outputs $\tilde{x}_y$ are paired up as expected. To mitigate this, CycleGAN [7] is proposed to introduce an additional cycle-reconstruction constraint for unpaired two-domain image translation. Specifically, the generated output is mapped back to the original image with an inverse translator, and the $\mathcal{L}_1$ loss is explicitly used as a loss

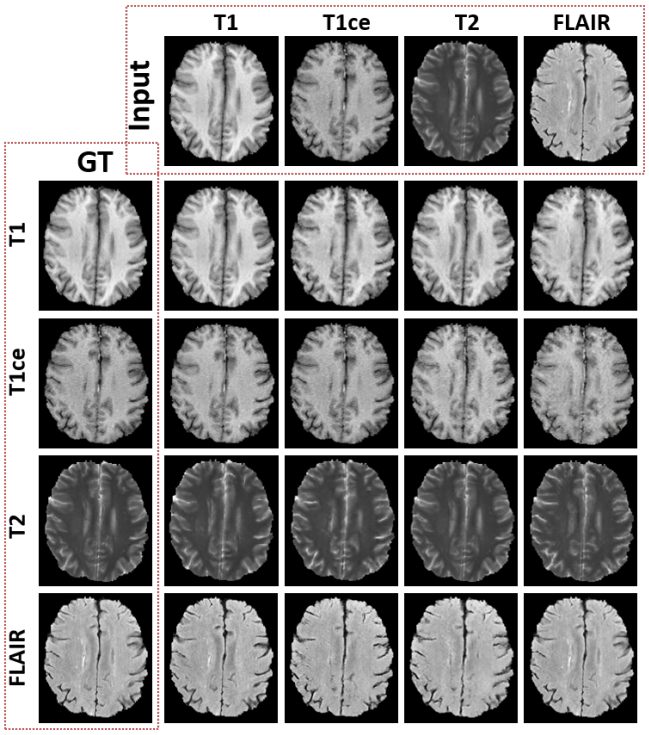

Fig. 2. Illustration of the results of our proposed framework. We use the first row as input, and configure four target modalities. The diagonal results are obtained using self-translation (i.e., $m_y = m_x$).

function to measure the similarity between the mapped back image and the original input. In this way, the shape structure can be better maintained. Note that both Pix2Pix [6] and CycleGAN [7] are used as the two-domain translators, and there are two inverse autoencoders to achieve the cycle reconstruction [7].

In addition, rather than configuring two autoencoders with inverse direction [7], we can simply recall the same autoencoder twice with a different conditional modality code in a feedforward processing. Specifically, the second time translation always uses the modality of original input sample m_x to achieve the reconstruction of x, given by

$$\min_{Enc,Dec} \mathcal{L}_1(\tilde{x}_x, x_x) = |\tilde{x}_x - x|, \tag{4}$$

where $\tilde{x}_x = Dec(Enc(\tilde{x}_y), m_x)$ is expected to reconstruct x.

To achieve the disentanglement of anatomical information and modality-specific factors without the anatomical label, we propose to enforce the similarity between $Enc(x_x)$ and $Enc(\tilde{x}_x)$ which are the two encoder outputs in a cycle, given by

$$\min_{Enc} \mathcal{L}_1^{disen} = |Enc(\tilde{x}_y) - Enc(x)|, \tag{5}$$

which explicitly requires that the paired co-registered two-modality images have the same feature map. Their shared factors can be the anatomical information [16], and have the difference between modality-specific imaging parameters

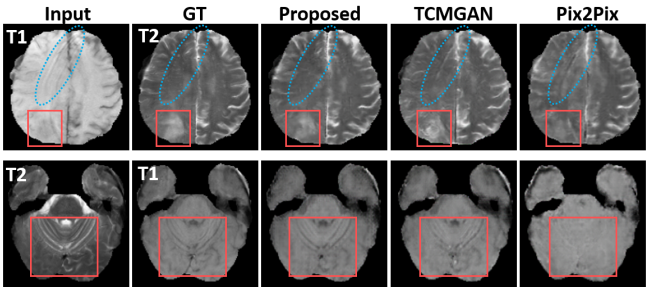

Fig. 3. Comparison of different methods for the T1-weighted and T2-weighted MR translation. GT indicates the ground truth x_y.

[16]. The feature vector is also required to be combined with the target modality label to reconstruct the target image, which encourages them to be independent and complementary to each other [17]. Therefore, the latent feature space can be anatomically related and modality-invariant (i.e., disentangled with a modality factor) [18, 17]. In addition, the feature-level similarity is also related to the perception loss [23], which enhances the textures.

2.2. Training Strategy

For simpler implementation, we reformulate the min-max terms to minimization only in a consistent manner. The objective of our conditional autoencoder and the adversarial cycle-reconstruction streams can be formulated as:

$$\min_{Enc,Dec} \mathcal{L}_1(\tilde{x}_y, x_y) + \alpha\mathcal{L}_1(\tilde{x}_x, x_x) + \beta\mathcal{L}_{adv} + \lambda_1\mathcal{L}_{mc}, \tag{6}$$

$$\min_{Dis} -\mathcal{L}_{adv} + \lambda_2\mathcal{L}_{mc}, \tag{7}$$

where α, β, λ_1, and λ_2 are the weighting hyperparameters. Enc and Dec minimize $\mathcal{L}_{adv}$, while Dis minimize $-\mathcal{L}_{adv}$ to play a round-based adversarial game to improve each other to find a saddle point. In practice, we sample the same number of x from each modality in training [6].

2.3. Testing Translation

After training, we can obtain the translation functions by assembling a subset of the subnetworks, i.e., Enc and Dec. Note that our translation can be agnostic to an input modality. Therefore, with an input sample x and a target modality m_y, we can generate its corresponding $\tilde{x}_y = Dec(Enc(x), m_y)$. With generated images with a target modality, we concatenate them for further tumor segmentation [10, 8].

3. RESULTS AND DISCUSSION

We evaluated our framework on the BraST'18 multimodal brain tumor database [9], which contains a total of 285 subjects with four MRI modalities: T1-weighted, T1ce, T2-weighted, and FLAIR MRI, with the size of 240×240×155. The intensity of slices was linearly scaled to $[-1, 1]$ as in [10, 5], which was then processed by 2D networks. The axial slices with less than 2,000 pixels in the brain area were filtered out as in [10].

Table 1. Numerical comparisons of four methods in testing

Methods	L1 ↓	SSIM ↑	PSNR ↑	IS ↑
12×Pix2Pix [6]	171.3±0.4	0.9206±0.0013	24.12±0.02	15.65±0.16
4×TCMGAN [10]	168.6±0.2	0.9413±0.0010	24.87±0.01	16.73±0.15
Proposed-$\mathcal{L}_1^{disen}$	159.8±0.3	0.9594±0.0016	25.21±0.02	18.10±0.13
Proposed	**157.2**±0.2	**0.9625**±0.0012	**25.92**±0.01	**18.54**±0.15

For a fair comparison, we followed [10] to use 100 subjects for the training translator, 85 subjects for testing, and 100 subjects for a training segmentor. We adopted the same backbone for Enc, Dec, and Dis for all comparisons [10]. In addition, the reimplemented Pix2Pix [6] was used as the two-modality transfer baseline model. In order to align the absolute value of each loss, we set weights $\alpha = 1$, $\beta = 0.5$, $\lambda_1 = 1$, and $\lambda_2 = 1$. We used Adam optimizer with a batch-size of 64 for 100 epochs training. The learning rate was set at $lr_{Enc,Dec} = 1\text{e}{-3}$ and $lr_{Dis} = 1\text{e}{-4}$ and the momentum was set at 0.5. Our framework was implemented using PyTorch. The training on an NVIDIA V100 GPU took about 8 hours. In practice, translating one test image with our unified Enc and Dec only took about 0.1 seconds.

3.1. Qualitative Evaluations

In Fig. 2, we illustrated the multimodal generation results of 12 cross-modality translation tasks and 4 self-reconstruction tasks. The proposed framework successfully synthesized any modality by simply configuring the target modality, which is consistent with the target ground truth MR images. We note the self-supervision was not used for the subsequent segmentation, but was used for checking the image generation quality in our implementation.

The qualitative comparisons with the 1-to-1 translator Pix2Pix [6] and 1-to-3 translator [10] are shown in Fig. 3 along with our proposed framework. The proposed framework was able to generate visually pleasing results with better shape and structure consistency when visually assessed. From the red box in the first row, we can see that the tumor area was better maintained with the help of the cycle-constraint compared with [10] which uses the additional tumor-consistent loss. Also, the artifact shown in the T1-weighted MR image (i.e., stripes indicated by the blue circle) yielded similar stripes as shown in the T2-weighted MR image with TCMGAN and Pix2Pix. Our disentangled encoder was able to eliminate the artifact and enforce the latent representation following the distribution of normal MR images.

3.2. Quantitative Evaluations

The synthesized images were expected to have realistic-looking textures, and to be structurally coherent with its corresponding ground truth images x_y. For quantitative evaluation, we adopted the widely used metrics including mean L1 error, structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), and inception score (IS).

Table 1 lists numerical comparisons between the proposed framework, Pix2Pix [6], and TCMGAM [10] for the 85 testing subjects using the BraTS'18 database. Of note, proposed-$\mathcal{L}_1^{disen}$ indicates the proposed model without the disentanglement constraint $\mathcal{L}_1^{disen}$. The proposed unified conditional disentanglement framework outperformed the other comparison methods w.r.t. these metrics, and the performance with the proposed framework was better than the proposed framework without the $\mathcal{L}_1^{disen}$.

Table 2. Comparisons of the segmentation accuracy

Methods	DICE
12×Pix2Pix[6]	0.7436±0.0017
4×TCMGAN[10]	0.7673±0.0014
Proposed-$\mathcal{L}_1^{disen}$	0.7791±0.0013
Proposed	**0.7862**±0.0015
Original 4 Modalities	0.8142±0.0012

3.3. Tumor Segmentation Results

In Table 2, we followed [10] to use the synthesized images by different methods to boost the tumor segmentation accuracy. Specifically, we sampled a slice with any modality in the testing data, and used our unified translation framework to generate its complementary three modalities [10]. Then, we concatenated the real slice and its generated three modalities as input to the segmentor. We note that only using the additional generated complementary slices can achieve improvements over only using one real slice [10]. For example, the DICE score was 0.7404 using only T2-weighted MRI for segmentation. We also computed the DICE score using the entire four real modalities, which served as an "upper bound".

The proposed unified conditional disentanglement framework yielded better segmentation performance than the baseline model Pix2Pix [6] and TCMGAN [10]. In addition, the DICE score of our conditional disentanglement framework was close to the upper bound which was computed using four real modalities. It was seen from our experiments that using all of the pairs in our training and the use of the cycle-constraint provided more accurate tumor shape recovery, thus leading to the better segmentation results.

4. CONCLUSION

This work presented a unified conditional disentanglement framework for co-registered multimodal translation based on a single set of target modality conditioned autoencoder and multi-task discriminator. The encoder is learned to extract the disentangled anatomical information by enforcing the consistency of two co-registered images with different modalities. The autoencoder is simply recalled twice to form a circular processing flow to enforce the cycle-constraint. Our framework is scalable to many modalities and effective to utilize the entire paired training data. In addition, our framework demonstrated superior performance on the tumor segmentation task over the compared methods using the generated images.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by BraTS'18. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

This work is partially supported by NIH R01DC018511, R01DE027989, and P41EB022544.

7. REFERENCES

[1] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, et al., "Medgan: Medical image translation using gans," *Computerized Medical Imaging and Graphics*, vol. 79, pp. 101684, 2020.

[2] X. Liu, F. Xing, C. Yang, C.C.-Jay Kuo, G. El Fakhri, and J. Woo, "Symmetric-constrained irregular structure inpainting for brain mri registration with tumor pathology," in *MICCAI BrainLes*, 2020.

[3] Y. Chang, G. C Sharp, Q. Li, H. A. Shih, G. El Fakhri, J. Ra, and J. Woo, "Subject-specific brain tumor growth modelling via an efficient bayesian inference framework," in *SPIE Medical Imaging*, 2018, p. 105742I.

[4] S. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE TMI*, pp. 2375–2388, 2019.

[5] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis," *IEEE TMI*, vol. 38, no. 7, pp. 1750–1762, 2019.

[6] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[7] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[8] Y. Qu, C. Deng, W. Su, Y. Wang, Y. Lu, and Z. Chen, "Multimodal brain MRI translation focused on lesions," in *ICMLC*, 2020, pp. 352–359.

[9] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE TMI*, 2014.

[10] B. Xin, Y. Hu, Y. Zheng, and H. Liao, "Multi-modality generative adversarial networks with tumor consistency loss for brain mr image synthesis," in *ISBI*, 2020.

[11] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018, pp. 8789–8797.

[12] X. Huang, M. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018, pp. 172–189.

[13] X. Yu, X. Cai, Z. Ying, T. Li, and G. Li, "Singlegan: Image-to-image translation by a single-generator network using multiple generative adversarial learning," in *ACCV*, 2018, pp. 341–356.

[14] W. Yuan, J. Wei, J. Wang, Q. Ma, and T. Tasdizen, "Unified attentional generative adversarial network for brain tumor segmentation from multimodal unpaired images," in *MICCAI*, 2019.

[15] M. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. Gotway, Y. Bengio, and J. Liang, "Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization," in *ICCV*, 2019, pp. 191–200.

[16] B. Dewey, L. Zuo, A. Carass, Y. He, Y. Liu, E. Mowry, S. Newsome, J. Oh, P. Calabresi, and J. L. Prince, "A disentangled latent space for cross-site MRI harmonization," in *MICCAI*, 2020, pp. 720–729.

[17] X. Liu, S. Li, L. Kong, W. Xie, P. Jia, J. You, and BVK Kumar, "Feature-level frankenstein: Eliminating variations for discriminative recognition," in *CVPR*, 2019.

[18] M. Mathieu, J. Jake Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *NeurIPS*, 2016, pp. 5040–5048.

[19] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NeurIPS*, 2017.

[20] A. Larsen, S. Sønderby, and H. Larochelle, "Autoencoding beyond pixels using a learned similarity metric," *ICML*, 2016.

[21] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *NeurIPS*, 2016, pp. 4743–4751.

[22] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *ICML*, 2017.

[23] X. Liu, BVK Kumar, P. Jia, and J. You, "Hard negative generation for identity-disentangled facial expression recognition," *Pattern Recognition*, 2019.