Context-Gan: Controllable Context Image Generation Using Gans

Marc-Adrien Hostin, Vladimir Sivtsov, Shahram Attarian, David Bendahan, Marc-Emmanuel Bellemare

CONTEXT-GAN: CONTROLLABLE CONTEXT IMAGE GENERATION USING GANS

Marc-Adrien Hostin[1,2], Vladimir Sivtsov[1], Shahram Attarian[3], David Bendahan[1], Marc-Emmanuel Bellemare[2]

[1]Aix Marseille Univ, CNRS, CRMBM, UMR 7339, Marseille, France
[2]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[3]Reference Center for Neuromuscular Diseases and ALS, La Timone, Aix Marseille Univ, Marseille, France

ABSTRACT

We propose an enhancement to label-to-image GANs. Based on a Pix2Pix architecture, ConText-GAN allows generating images in a controlled way. Given a feature map as input, ConText-GAN can generate images with a specified layout and label content. As an application, ConText-GAN is used to perform a more realistic than usual data augmentation from an MRI dataset. We show the validity of the generated images with respect to the input feature maps. The relevance of the approach is demonstrated by the improvement of the segmentation result following a data augmentation performed with ConText-GAN compared to classical methods. A practical application is presented in the challenging context of U-Net segmentation of MRI of fat infiltrated muscles.

***Index Terms*—** GAN, data augmentation, label-to-image, semantic image synthesis, MRI segmentation

1. INTRODUCTION

Data augmentation using generative adversarial networks (GAN) is a major focus for improving the training of neural networks applied to medical image analysis [1, 2]. Learning methods have been successful in many tasks such as disease classification or organ segmentation. However, the accuracy of data-driven methods is dependent on the amount of data available, which can be scarce, especially in the medical field.

Scientists have compensated for the lack of data by creating annotated synthetic images. The most naive approaches use geometric transformations (rotation, translation, elastic deformation) to synthesize data. These techniques have become established standards and can be useful when it comes to making a model robust to variations in orientation, position or shape [3]. However, the generated data may be too close to the original distribution, and may not provide any real benefit.

To solve this problem and improve the generalization ability of learning methods, we have been looking for a way to create realistic and more diverse data. Thanks to their ability to produce new images, GANs could be good candidates. Among the variety of GANs, we focused on the family of label-to-image translation methods. In addition to the creation of annotated data, these methods allow constraining the position of certain features of the image, *i.e.* organs of interest. Besides, to avoid the problem of class imbalance, we have to make sure that the resulting images are evenly distributed among the different classes considered (*e.g.* control/patient).

As a solution, we designed the ConText-GAN, a network capable of generating images respecting a spatial context specified by the user. Assuming that the spatial context (in our case a texture descriptor for each organ) is representative of the different classes in the image generation, our method allows creating a dataset with as many examples of each desired class. As an application, we propose to apply our method to the generation of MRIs of thighs of patients with neuromuscular diseases (NMD). NMDs affect tissues by progressively replacing muscle with fat, which is called fatty infiltration. The infiltration reduces the visibility of muscle contours, which can prevent them from being detected [4]. As the diseases concerned are rare and the number of patients limited, being able to synthesize images with the desired infiltrate texture is of great interest for deep learning segmentation. To our knowledge, only two studies have tested a GAN-based approach to perform data augmentation in this domain, using noise-to-image GAN [5] and image-to-image translation [6], but neither proposed to control the diversity of synthesized muscles. The few methods that have tried to control the texture of synthetic images have used a texture patch as context information [7]. However, the use of patches limits texture creation to the patches available in the database. Furthermore, it is difficult to evaluate the similarity between the generated textures and the input patches, which raises a major problem of GANs, namely the evaluation of their performance [1]. We replace the patch with spatially distributed statistical features on a map. This feature map allows us to both generate new contexts as input to our GAN and to evaluate the context rendering on the synthetic images.

In this study, we are proposing to use texture descriptors as a prior context to control the localized texture diversity provided by ConText-GAN. Our method was evaluated based on image quality criteria (L1, structural similarity index measure [SSIM], peak signal-to-noise ratio [PSNR]) and texture controllability (custom metric). Finally, a 2D U-Net [8] was trained for muscle segmentation from the augmented data, and the resulting segmentations were evaluated accord-

ing to Dice (DSC), average surface distance (ASD) and Hausdorff Distance (HD). Our main contributions are as follows:

- **ConText-GAN:** A ConText-GAN for generating images from context prior was developed to address the class imbalance problem in data augmentation.
- **Muscle MRI:** This work shows that fat infiltrated muscles can be generated in a controlled fashion to create a balanced dataset.

2. METHODS

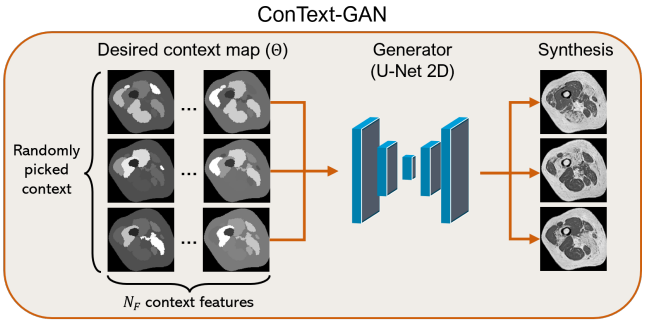

Fig. 1. From the same topology (muscle position), and 3 different context maps, ConText-GAN synthesis of 3 MRI of muscular thighs.

This section describes the principle of ConText-GAN in a general framework and further describes the database used in this study along with the evaluation process.

2.1. ConText-GAN

Among label-to-image models, we chose Pix2Pix [9] as a reference to prove the validity of our proposal, but newer models such as OASIS [10] could be used in the future to improve the quality of the synthesis. It consists in training a generator G (U-Net) and a discriminator D (PatchGAN [9]). Let x be the label domain and y the image domain, G aims at translating a combination of x and a random noise vector z to y and create a synthetic image ($\in \hat{y} = G(x)$). D takes as input a label from x and an image, its purpose is to differentiate between real images ($\in y$) and synthetic images ($\in \hat{y}$). G and D are trained in an adversarial manner, where G is trained to fool D, while D is trained to improve its discriminative capability. The loss function of a cGAN ($\mathcal{L}_{cGAN}$ (1)) is minimized by G against D which tries to maximize it. Pix2Pix adds a distance loss term $\mathcal{L}_D$ with a coefficient λ (set to 100), to take into account the similarity between the produced and the real images. Resulting loss function $\mathcal{L}$ is expressed in (2):

$$\mathcal{L}_{cGAN} = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z)))] \quad (1)$$

$$\mathcal{L} = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_D \quad (2)$$

Our main goal is to enable image context control by adding descriptors as input to the GAN, which we named ConText-GAN (Fig. 1). Let Θ be a $N_F \times h \times w$ tensor and i, j pixel coordinates with N_F being the number of context features, h and w the height and width of the input image. A context map would be defined by $\{\Theta_{i,j}\} = (\alpha^0_{i,j}, ..., \alpha^n_{i,j})$ where $\alpha^k_{i,j}$ is the value of the k^{th} context feature in pixel coordinates i, j. Let X be a segmentation mask, the synthetic image produced with ConText-GAN $\hat{Y}$ would be expressed as $\hat{Y} = G(\Theta, X)$. The addition of Θ at the input of the network questioned our choice of distance loss $\mathcal{L}_D$. We evaluated several $\mathcal{L}_D$ functions for training the generator: $\mathcal{L}_{L1}$ (3), the loss used in the original Pix2Pix paper [9]; $\mathcal{L}_R$ (4), which includes the $\mathcal{L}_{L1}$ as well as a perception-based loss function $\mathcal{L}_S$ using the structural similarity index measure (SSIM [11]) for the purpose of obtaining realistic images; $\mathcal{L}_{CT}$ (5), that contains $\mathcal{L}_R$ as well as a function $\mathcal{L}_T$ to constrain the context synthesis. $\mathcal{L}_T$ consists in measuring the L1 distance between the context descriptors measured respectively on the estimated image and the real image. Since the loss function must be differentiable to be optimized, the choice of descriptors must focus on a differentiable function, or a derivable estimate of the measure. In the end, we propose the following loss functions to be tested with ConText-GAN :

$$\mathcal{L}_{L1}(y, \hat{y}) = |y - \hat{y}| \quad (3)$$

$$\mathcal{L}_R(y, \hat{y}) = \mathcal{L}_{L1}(y, \hat{y}) + \mathcal{L}_S(y, \hat{y}) \quad (4)$$

$$\mathcal{L}_{CT}(y, \hat{y}) = \mathcal{L}_R(y, \hat{y}) + \mathcal{L}_T(y, \hat{y}) \quad (5)$$

2.2. Dataset

We used our own dataset to evaluate the ConText-GAN. The database consists of 170 acquisitions from 102 subjects, including 14 controls and 88 NMD patients. MRI scans were recorded at 1.5T (MAGNETOM Avanto, Siemens Healthineers, Erlangen, Germany) at the thigh level using a spine coil on the bottom and a flexible coil on top of the lower limb. One image set consisted of 2D T_1-weighted MRI (T_{1W}) (TR = 578 ms; TE = 11 ms; FA = 90°; bandwidth = 182 Hz/pixel; in-plane matrix size/voxel size = $320 \times 160/1.26 \times 1.26$ mm^2; 38 slices [slice thickness = 4.40 mm]; slice gap = 0.40 mm). The second set T_{1w} MRI acquisition featured : TR = 549 ms; TE = 11 ms; FA = 120°; bandwidth = 195 Hz/pixel; in-plane matrix size/voxel size = $320 \times 320/0.68 \times 0.68$ mm^2; 20 slices [slice thickness = 10.00 mm]; slice gap = 5.00 mm. In total, 3990 slices were annotated. 12 labels including adductor, bone, *biceps femoris*, *gracilis*, *rectus femoris*, *sartorius*, *semimembranous*, *semitendinosus*, extramuscular fat, *vastus intermedius*, *vastus lateralis*, and *vastus medialis*, have been segmented by different experts. The images were all resized to 256×256, and a gaussian blur was applied on the most resolved images to obtain the same resolution in the whole set.

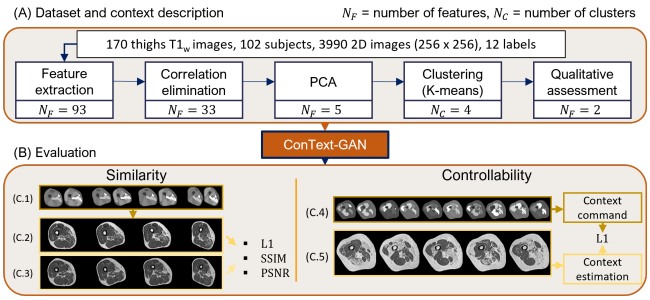

Fig. 2. A) Feature selection for context description. B) Evaluation process. C.1) Feature maps with same context, different topology. C.2), C.3) Generated and reference MRIs. C.4) Feature maps with same topology, different context. C.5) Generated MRIs from C.4). (Best zoomed)

2.3. Context description

The selection of the context depends on the desired data augmentation. Here, we sought to produce MRIs of thigh muscles at different levels of fat infiltration, to get a dataset with equivalent numbers of pathological and healthy muscles. Context selection was based on radiomics analysis, which consists in measuring numerous features on the images (93 in our case, using the pyradiomics library [12]), and then use statistical methods to identify the features most related to a clinical problem (*e.g.* fat infiltration). The process is described in Fig. 2.A). Pairwise correlation analysis was used to eliminate features that were too correlated with each other ($|p| > 0.8$ with p the Pearson coefficient between two features). A principal component analysis (PCA) on the remaining 33 descriptors identified 5 major components. One feature per component was selected by systematically choosing the most correlated with the component. The remaining components were selected for their ability to differentiate fatty infiltration textures. To assess the relevance of descriptors, the usual method is to analyze the correlation between descriptors and a clinical score of interest. In the NMD study, the only approved clinical score is the fat fraction [13] (or its qualitative version, the Mercuri score [14]), which represents the proportion of fat in muscle, *i.e.* the average of the intensities on the image. However, it was not relevant to choose these scores as clinical references, since the correlation study would have revealed a significant correlation only with the mean as a descriptor. In the absence of a clinical reference, we used our features to group our dataset into four clusters. It is easy to differentiate, by observation, four classes: healthy muscles, totally infiltrated muscles, and two intermediate classes corresponding to cases of mild and severe infiltration. Beyond four, the differences were too tenuous to be visible to the human eye. Five descriptors resulting from the PCA were used to perform k-means clustering. Based on a qualitative assessment of the clustering accuracy for each descriptor, two features were selected: Average intensity, which represents the level of whiteness of the muscle; and entropy, which represents the disorder in the texture, a measure of the dispersion of the fatty infiltrate.

2.4. Evaluation

Evaluation of GAN generated images is controversial because it often lacks quantitative scores [1]. However, in our case, it was possible to make a direct comparison between the generated images and the reference images. For this purpose, a test set with deformed image/label pairs, with the same elastic deformation, was generated, as shown in Fig. 2.B). Thus, it was guaranteed that the GAN had never been trained on these labels. Then, synthetic images were produced from the distorted labels and compared to corresponding reference images. Comparisons were performed using images from the baseline Pix2Pix (without texture control) trained with different $\mathcal{L}_D$ loss: B$_{L1}$ used $\mathcal{L}_{L1}$ and B$_R$ used $\mathcal{L}_R$. Similarly, we got 3 models from our ConText-GAN : CT$_{L1}$ was trained with $\mathcal{L}_{L1}$, whereas CT$_R$ used $\mathcal{L}_R$, and CT$_T$ used $\mathcal{L}_T$. All models were assessed using the same criteria. Two main criteria were evaluated with this process. First, the similarity between the generated and reference images was measured with the L1, SSIM and PSNR distance scores on every network. Second, to verify the controllability of the network, the context description map was measured on the images produced by the ConText-GAN and then compared with the context map used as input with an L1 loss. Finally, the diversity of the images produced was checked by ascertaining the percentage of muscles located in each of the four grouped clusters using k-means applied on the mean and entropy features computed on the predicted images.

2.5. Data augmentation for segmentation

As previously stated, data augmentation is typically used at the training step to enhance CNN generalization ability. We advocate that the generated images be in the image domain, as diverse and as distant as possible from the original distribution. Images generated by ConText-GAN should meet this condition thanks to its control capability to randomly assign shapes and textures to images. As application, we used our method with the training database taken from the T$_{1w}$ muscle MRI database to be segmented with CNNs. To ensure synthesized muscle texture diversity, the realistic augmentation induced was performed as follows: New labels were generated by random elastic deformation to provide new muscle shapes; Each label was associated with a texture descriptor (*e.g.* pair mean-entropy) randomly picked among those of the real dataset.

We compared the impact of data augmentation methods on the segmentation results of CNNs. We have trained various U-Net models against various training datasets. The core dataset depicted in section 2.2 being the same: N$_{aug}$ used it without any augmentation; A$_E$ used a dataset augmented with

elastic deformation; A_{P2P} used Pix2Pix augmentation; A_{CT} used ConText-GAN data augmentation done with CT_T.

3. RESULTS AND DISCUSSION

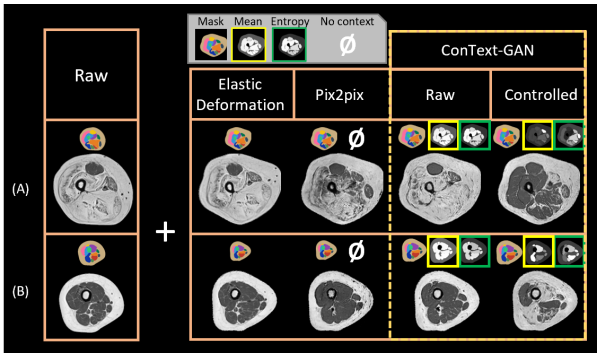

Fig. 3. Image augmentation with elastic deformation, Pix2Pix and ConText-GAN using raw context from existing image and randomly generated context on patient (A) and control (B).

The images produced by ConText-GAN seemed more realistic than those from the standard Pix2Pix, as seen in Fig. 3. On the right column, the creation of an image with the same topology and a different context shows the ability of the network to assign new muscle textures.

	L1	SSIM	PSNR
B_{L1}	4.481 ± 1.976	0.799 ± 0.068	20.466 ± 2.041
B_R	4.342 ± 2.001	0.804 ± 0.069	20.808 ± 2.192
CT_{L1}	4.096 ± 1.713	0.805 ± 0.073	21.260 ± 2.012
CT_R	4.083 ± 1.689	0.810 ± 0.071	21.364 ± 1.972
CT_T	**4.004 ± 1.758**	**0.818 ± 0.067**	**21.564 ± 2.152**

Table 1. Similarity between reference image and synthetic produced with Pix2Pix (B_{L1},B_R) and ConText-GAN (CT_{L1}, CT_R and CT_T).

Tab. 1 shows that the best results regarding the similarity of the images produced with the reference data were obtained with CT_T (L1: 4.00 ± 1.76, SSIM: 0.82 ± 0.06, PSNR: $21.56.0 \pm 2.15$). ConText-GAN can synthesize more realistic images than a standard cGAN like Pix2Pix. In addition, the feature map based control of ConText-GAN allowed to generate a balanced training database. A k-means clustering performed with the generated muscle texture features has provided a balanced class distribution of the samples in the four pathological muscle clusters that constitute our test base.

The results of the context controllability study (Tab. 2) showed that the error on the mean was lowest for CT_{L1} (3.02 ± 1.13), but the relative errors with CT_R (14%) and CT_T (7%) were quite close. For the entropy measure, the

	Mean	Entropy
CT_{L1}	**3.023 ± 1.131**	0.419 ± 0.131
CT_R	3.512 ± 1.232	0.400 ± 0.114
CT_T	3.265 ± 0.933	**0.264 ± 0.086**

Table 2. Report of L1 difference between the texture features measured on reference and synthetic images produced by ConText-GAN models CT_{L1}, CT_R and CT_T respectively trained with $\mathcal{L}_{L1}$, $\mathcal{L}_R$, $\mathcal{L}_T$

results obtained with CT_T (0.26 ± 0.09) were significantly better than those given by CT_{L1} (37%) and CT_R (34%). This can be interpreted stating that the loss $L1$ is sufficient for a feature as simple as the average pixel value, whereas for a higher order feature it is more rewarding to include an estimate in the loss function.

	Multiplication factor	DSC	ASD (mm)	HD (mm)
N_{aug}	1	0.862 ± 0.145	1.580 ± 2.639	8.354 ± 13.265
A_{P2P}	2	0.862 ± 0.143	1.468 ± 2.013	7.510 ± 11.940
A_{P2P}	4	0.853 ± 0.158	1.675 ± 2.659	8.482 ± 13.387
A_E	2	0.866 ± 0.140	1.451 ± 2.274	7.326 ± 11.558
A_E	4	0.870 ± 0.135	1.371 ± 1.958	6.862 ± 10.842
A_{CT}	2	0.873 ± 0.123	1.273 ± 1.658	6.283 ± 9.927
A_{CT}	4	**0.875 ± 0.115**	**1.243 ± 1.382**	**6.188 ± 9.073**

Table 3. Impact of training dataset on the segmentation result of U-Net. N_{aug}: raw dataset, A_E: elastic deformation, A_{P2P}: pix2pix, A_{CT}: ConText-GAN.

Best segmentation result was achieved with the ConText-GAN augmented training dataset: A_{CT} with an increase factor of 4 (DSC=0.875 ± 0.115, ASD=1.243 ± 1.382 mm, HD=6.188 ± 9.073 mm). One can note that the results obtained with the A_E and A_{CT} gave better results compared to the benchmark N_{aug}. A_{CT} led to a lower error than A_E (*c.f.* Tab. 3). The balanced database obtained thanks to the ConText-GAN allowed to improve the accuracy of the segmentation.

4. CONCLUSION

We have developed an improvement to label-to-image GANs. ConText-GAN generates realistic images respecting content constraints imposed with feature maps. As an example, we have shown how useful this capability can be for performing realistic data augmentation in medical imaging.

The addition of the feature map did not overly complicate the use of GAN, and while we started with medical textures, we could try ConText-GAN in other areas. Obviously, the choice of texture parameters may also evolve for different cases and, apart from the problem of the differentiability of the measurements to fill the loss function, the results should be transferable.

5. ACKNOWLEDGMENTS

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

6. COMPLIANCE WITH ETHICAL STANDARDS

The study was approved by the local human research committee and was conducted in conformity with the Declaration of Helsinki (version October 2013) and the Medical Research Involving Human Subjects Act. Prior written informed consent was obtained from all subjects.

7. REFERENCES

[1] Vera Sorin, Yiftach Barash, Eli Konen, and Eyal Klang, "Creating artificial images for radiology applications using generative adversarial networks (gans)–a systematic review," *Academic radiology*, vol. 27, no. 8, pp. 1175–1185, 2020.

[2] Yizhou Chen, Xu-Hua Yang, Zihan Wei, Ali Asghar Heidari, Nenggan Zheng, Zhicheng Li, Huiling Chen, Haigen Hu, Qianwei Zhou, and Qiu Guan, "Generative adversarial networks in medical image augmentation: a review," *Computers in Biology and Medicine*, p. 105382, 2022.

[3] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth, "A review of medical image data augmentation techniques for deep learning applications," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, no. 5, pp. 545–563, 2021.

[4] Augustin C Ogier, Marc-Adrien Hostin, Marc-Emmanuel Bellemare, and David Bendahan, "Overview of mr image segmentation strategies in neuromuscular disorders," *Frontiers in Neurology*, vol. 12, pp. 625308, 2021.

[5] Michael Gadermayr, Kexin Li, Madlaine Müller, Daniel Truhn, Nils Krämer, Dorit Merhof, and Burkhard Gess, "Domain-specific data augmentation for segmenting mr images of fatty infiltrated human thighs with neural networks," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 6, pp. 1676–1683, 2019.

[6] Michael Gadermayr, Lotte Heckmann, Kexin Li, Friederike Bähr, Madlaine Müller, Daniel Truhn, Dorit Merhof, and Burkhard Gess, "Image-to-image translation for simplified mri muscle segmentation," *Frontiers in Radiology*, p. 3, 2021.

[7] Dario Augusto Borges Oliveira, "Controllable skin lesion synthesis using texture patches, bézier curves and conditional gans," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1798–1802.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[10] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva, "Oasis: Only adversarial supervision for semantic image synthesis," *International Journal of Computer Vision*, vol. 130, no. 12, pp. 2903–2923, 2022.

[11] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[13] Jasper M Morrow, Christopher DJ Sinclair, Arne Fischmann, Pedro M Machado, Mary M Reilly, Tarek A Yousry, John S Thornton, and Michael G Hanna, "Mri biomarker assessment of neuromuscular disease progression: a prospective observational cohort study," *The Lancet Neurology*, vol. 15, no. 1, pp. 65–77, 2016.

[14] Eugenio Mercuri, Beril Talim, Behzad Moghadaszadeh, Nathalie Petit, Martin Brockington, Serena Counsell, Pascale Guicheney, Francesco Muntoni, and Luciano Merlini, "Clinical and imaging findings in six cases of congenital muscular dystrophy with rigid spine syndrome linked to chromosome 1p (rsmd1)," *Neuromuscular Disorders*, vol. 12, no. 7-8, pp. 631–638, 2002.