MetH: A family of high-resolution and variable-shape image challenges

Ferran Parés[1] and **Dario Garcia-Gasulla**[2] and **Harald Servat**[3] and **Jesús Labarta**[4] and **Eduard Ayguadé**[5]

Abstract. High-resolution and variable-shape images have not yet been properly addressed by the AI community. The approach of down-sampling data often used with convolutional neural networks is sub-optimal for many tasks, and has too many drawbacks to be considered a sustainable alternative. In sight of the increasing importance of problems that can benefit from exploiting high-resolution (HR) and variable-shape, and with the goal of promoting research in that direction, we introduce a new family of datasets (MetH). The four proposed problems include two image classification, one image regression and one super resolution task. Each of these datasets contains thousands of art pieces captured by HR and variable-shape images, labeled by experts at the Metropolitan Museum of Art. We perform an analysis, which shows how the proposed tasks go well beyond current public alternatives in both pixel size and aspect ratio variance. At the same time, the performance obtained by popular architectures on these tasks shows that there is ample room for improvement. To wrap up the relevance of the contribution we review the fields, both in AI and high-performance computing, that could benefit from the proposed challenges.

1 INTRODUCTION

Challenging problems to solve is what drives AI research and pushes the field and its applications forward. A prime example of that is the ImageNet dataset together with the corresponding ILSVRC challenge [43]. The popularization of this competition revitalized the Neural Networks field, particularly in the context of image processing. The outstanding performance of deep neural networks models in the ILSVRC challenge caught the attention of AI researchers and practitioners, who quickly acknowledged the potential behind the combination of deep nets and large sets of data. As a result, the popularity of the field exploded.

The ImageNet dataset provided an appealing challenge to lure researchers, who in turn, developed and tested many new ideas on it. Some of these ideas became powerful principles for the current deep learning (DL) field: Inception blocks [49], ResNet architecture along its shortcuts [22], Dropout [48] and ReLU [37], among others. This resulted in remarkable achievements in an extraordinary short time. Nowadays, the relevance of the ImageNet image classification challenge has mostly vanished, as its considered to be a solved problem for the AI community. By of 2019, 98.2% top-5 accuracy [56] was achieved, while human top-5 classification error is between 12% and 5% as stated in the original ILSVRC work [43]. ILSVRC 2017 was the last edition of the challenge, as announced by the organizers [2].

Figure 1. Subset of images from the official Met dataset. Notice the variation in shape among images.

The necessity of new challenges to push the AI field forward is constant, particularly considering the current pace of research. Examples of these are deep fake video detection [3, 4, 5] or motion recognition in videos [28, 53]. The first task aims to automatically detect videos where facial traits are artificially modified to change the appearance of a person. This is an important task at the moment as recent deep fake methods enable the creation of applications capable of directly manipulating society. The second task targets the identification and classification of a given set of corporal gestures. This research has many real world applications, in fields like assistive technologies, Human-computer interaction and law enforcement.

In this paper we propose a set of visual challenges, focused on two aspects of image related tasks that have been overlooked so far: High-resolution (HR) and variable-shape images. In the context of this paper, we consider HR datasets those composed by images equal or larger than 500x500 pixels. This definition is based on the fact that a vast majority of current image recording devices are capturing images at least at that resolution, which makes anything below that low-resolution.

HR is fundamental for AI research, as many present and future visual challenges can be better solved by having HR insight into the data. Two good examples of this are medical imaging and autonomous driving. Many visual challenges in the medical domain are based on HR images, and solving them properly require both, attention to detail and understanding of large structures. In domains like breast cancer detection, the benefit of exploiting the highest possible resolution has already been highlighted [17, 33]. In autonomous driving, using HR images entails detection at further distances, which has enormous safety implications. Current solutions already use images

[1] Barcelona Supercomputing Center (BSC), email: ferran.pares@bsc.es
[2] Barcelona Supercomputing Center (BSC), email: dario.garcia@bsc.es
[3] Intel Corporation Iberia S.A., email: harald.servat@intel.com
[4] Barcelona Supercomputing Center (BSC), Universitat Politcnica de Catalunya (UPC), email: jesus.labarta@bsc.es
[5] Barcelona Supercomputing Center (BSC), Universitat Politcnica de Catalunya (UPC), email: eduard.ayguade@bsc.es

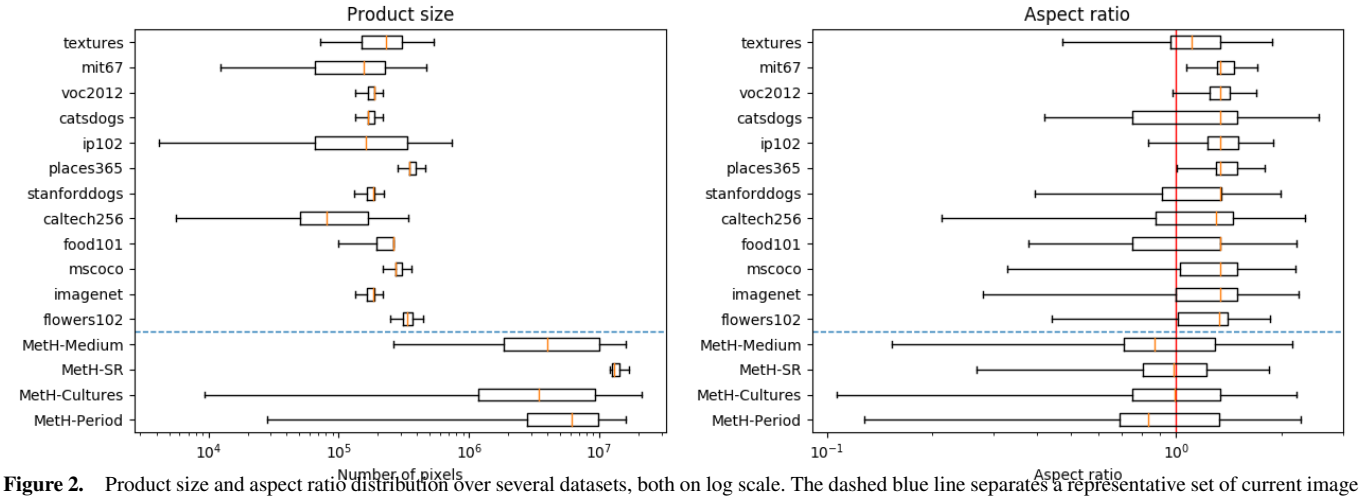

Figure 2. Product size and aspect ratio distribution over several datasets, both on log scale. The dashed blue line separates a representative set of current image classification datasets, and the MetH datasets introduced here. The vertical red line at aspect ratio 1.0 shows the border between portrait (left side) and landscape (right side) images.

that fall within our definition of HR [11, 51].

The second image feature we analyze in this paper is variable-shape. There are several datasets in the field already containing images of different shape. However, datasets of this kind often have relatively small variations in aspect ratio, as discussed in §2. The AI community should start working on datasets with a wider image shape variety, as these will become more common. The prime example of that tendency are crowd-sourced datasets, like the Open Images dataset [30]. Building datasets by combining multiple sources saves time and effort, but implies that images are taken using a variety of recording devices, which will most likely not have the same image resolution or even shape. Additionally, the orientation of the device when taking the picture can also imply changes in shape (*i.e.*, landscape or portrait). Most of the current approaches interpolate all images to ensure that they have the same shape, avoiding to deal with variable-shape. However, this approach has a clear drawback: the loss and/or deformation of information. This directly affects the models performance [18].

To promote and empower research to tackle these issues, in this work we introduce a series of datasets containing images of high-resolution and variable-shape. The associated tasks are image classification (x2), regression and image super resolution. All images contained in these datasets are obtained from the Metropolitan Museum of Art (Met). The Met released the data under Creative Common Zero (CC0) license [6].

2 RELATED WORK

There are many visual challenge datasets in the current literature. There are however, very few with datasets containing images larger than 500x500 pixels, and with a significant variance along aspect ratio. To illustrate that point we analyze 12 popular datasets which satisfy the three following conditions:

- The dataset must be publicly available.
- The dataset labels must be reliable.
- The dataset must have at least 100 instances per class.

The first condition is self-explanatory. The second one excludes all datasets that contain labels not validated by humans or that have been crowd-labeled, as these contain a significant amount of noise. The third enforces a minimum number of instances, as we consider these necessary for thorough research experimentation. We were nonetheless flexible in this regard, as some datasets analized here have a majority of classes with more than 100 instances but also a few classes with less. The final set of analyzed datasets is: ImageNet 2012 [43], Food101 [9], IP102 [55], Places365 [60], Mit67 [42], Flower102 [38], CatsDogs [41], StanfordDogs [26], Textures [12], Caltech256 [20], Microsoft COCO [32] and Pascal VOC 2012 [13].

We analyze the product size (*i.e.*, width multiplied by height) and the aspect ratio (*i.e.*, width divided by height) distributions of each dataset. For the three datasets larger than 100,000 images (ImageNet 2012, Places365 and Microsoft COCO) we take a sample of that size. Distributions for all 12 datasets can be seen in Figure 2.

In terms of number of pixels (left plot), current image classification datasets do not contain images with more than 1 megapixel (MP). This indicates a bias in current research, particularly considering that currently popular resolutions are much larger than that. In contrast, the average image sizes in our datasets MetH-Cultures, MetH-Medium, MetH-Period and MetH-SR are 5.1, 5.5, 6.3 and 13.6 MP, respectively. Obviously, there are datasets today with images larger than 1 MP, however, these are typically either private, unreliability labeled [30], or have too few instances per class [14].

Regarding aspect ratio, the right plot of Figure 2 shows how the majority of images found in current datasets are landscape with minimal variations. For example, in the ImageNet dataset [43] most images have both width and height within the 400 and 600 pixel range. Some datasets do contain a significant amount of portrait images, such as the Food101, CatsDogs and Caltech256 datasets. However, even in these cases, their aspect ratio distribution is clearly skewed towards landscape images (notice that the median is quite close to the third quartile on all three cases).

The bottom side of the plots in Figure 2 shows the 4 datasets proposed in this paper. These datasets are introduced and explained in the following section. Notice that our datasets contain images of higher resolution. In fact, all images in the Q1-Q3 interval (the boxes of Figure 2) of the MetH datasets are bigger than the largest image found on all analyzed datasets. Furthermore, the mean image size for the MetH datasets is at least one order of magnitude larger on all

Figure 3. Samples of MetH-Medium. Top row shows samples of *hard-paste porcelain*, *soft-paste porcelain* and *generic porcelain*. Bottom row shows examples of *limestone* and *generic stone*. This shows the difficulty of the problem, even for humans.

cases, and even more for the MetH-SR dataset. Regarding the aspect ratio, our proposed datasets have a balanced distribution, containing more or less as many portrait as landscape images. This distribution is also rather wide, which means MetH datasets contain both very wide and very tall images. Also MetH-Period and MetH-Medium contain slightly more portrait than landscape images. These properties characterize our datasets and make them relevant in the current context.

Finally, let us discuss current super resolution datasets, as none of the datasets analyzed in this section are specific for this task. Among the most popular super resolution datasets, we can find challenges with either HR [50] or variable-shape properties [7, 8, 24, 35, 59], but not both. Moreover, when HR is present [50], this is limited to 2K resolution and 1,000 samples, with a maximum target scale factor of x4. Our proposed dataset is larger in number of samples, in resolution and in target scaling factor. It remains to be seen how current state-of-the-art will perform in this new and demanding setting.

Dataset 1	Dataset 2	Images %
MetH-Medium	MetH-Cultures	7.66%
MetH-Medium	MetH-Period	20.81%
MetH-Medium	MetH-SR	8.82%
MetH-Cultures	MetH-Medium	14.09%
MetH-Cultures	MetH-Period	8.64%
MetH-Cultures	MetH-SR	10.72%
MetH-Period	MetH-Cultures	4.82%
MetH-Period	MetH-Medium	21.38%
MetH-Period	MetH-SR	8.06%
MetH-SR	MetH-Cultures	8.04%
MetH-SR	MetH-Medium	12.16%
MetH-SR	MetH-Period	10.82%

Table 1. Percentage of images in dataset 1 also found within dataset 2

3 MetH DATASETS

The Metropolitan Museum of art has been photographing its museum pieces of art in HR for the records. In 2017, the museum decided to released some of their photographs along with associated metadata under the Creative Commons Zero (CC0) license [6]. The metadata is available as a CSV file posted on the Met official GitHub repository. We downloaded all the images with metadata by crawling the Met website on March 2019 [39].

Figure 4. Samples of MetH-Cultures. All samples belong to different cultures: British, Chinese, Italian, German, French, Japanese and Spanish. These set of porcelain pieces exemplifies the subtle differences between cultures.

Based on the Met images, we define four datasets, all of them including images of high-resolution and variable-shape: Two datasets of image classification representing different problematics, the MetH-Medium and the MetH-Culture. One of regression based on the visual aspect of the images, the MetH-Period. And one of super resolution, the MetH-SR. Our datasets share some images between them, check the specific intersections in Table 1.

All four datasets are publicly available [6]. Each dataset contains a folder with all the images and a CSV file with metadata for all images in the folder. The metadata includes the target of the problem and the train, validation and test splits. All datasets names are prefixed with *MetH*, a combination of Met and the name of the research group in which this contribution was developed (HPAI).

3.1 Medium

The Medium dataset (MetH-Medium) targets the identification of the medium a piece of art is made of (*i.e.*, the main material of the piece) based on its visual appearance. The dataset has 23 different classes, including Gold, Silver, Woodcut, Limestone or Silk. All these classes are balanced on the number of instances, each containing 1,000 images for training, 100 images for validation and 100 images for test. By aggregation, the entire dataset contains a total of 27,600 images: 23,000 images for training and 2,300 images for both validation and test.

In general, this seems like a dataset oriented towards the detection of small-scale patterns. HR images can help a lot in that regard, providing small detail on the art pieces. That being said, we cannot discard that larger patterns are useful for classification, since different materials are used differently (*e.g.*, totems are not made of silk). Furthermore, there are 10 classes within the dataset which are quite hard to discriminate, even for humans. Examples of these are shown in Figure 3. Having such a fine-grained subset of classes guarantees the challenging nature of the dataset, and forces any potential solution to exploit patterns found at every possible level of detail.

3.2 Cultures

The Cultures dataset (MetH-Cultures) targets the identification of the culture in which a piece of art was created. This dataset considers pieces of art belonging to 15 different cultures, including Roman, British, Coptic, Etruscan or Greek. All classes contain the same number of instances: 800 instances for training, 100 instance for validation and another 100 for test. The total amount of instances is 12,000 for training and 1,500 for validation and test.

[6] `https://hpai.bsc.es/meth-datasets`

This dataset is a good candidate for exploiting HR in images. It requires identifying the object while paying attention to details that help discriminate cultures. To illustrate that point, in Figure 4 we show a variety of porcelain pieces of art made by different cultures. Some of the porcelain pieces can be partially discriminated based on its shape (*e.g.*, bowls are typical of Asian cultures), while others require paying attention to the small drawings decorating the art (*e.g.*, some trees or flowers may be specific of a certain culture).

3.3 Period

The Period dataset (MetH-Period) is a regression problem where the goal is to predict the year in which a piece of art was crafted. The data released by the Met contains the exact year of creation for some pieces, and a range of years for others (when the exact year could not be established). In order to reduce noise in the proposed task, we include only art pieces for which the exact year of creation is known. The final range of years for this task goes from year 1500 to year 1900, capturing pieces from the very beginning of the Modern Age to the middle of the recent Contemporary Age. The training and validation sets are balanced when aggregating images in buckets of 50 years. The distribution of samples aggregated in this manner can be seen in Figure 5. This dataset has a total amount of 37,762 instances distributed into 8,000 training, 2,400 for validation and 27,362 for testing. This accounts for an average of 20 samples per year in the train set.

Over the considered time frame, art tendencies have varied significantly. In Europe, art evolved through many periods, including Rennaissance, Neo-Classicism, Romanticism, Realism, Impressionism and early Modern. In China through the Ming and Qing periods. In Japan through the Momoyama, Edo and Meiji Restoration periods. In this context, it seems clear that learning to solve the MetH-Period tasks requires learning at least some basic properties of this evolutionary process, paying attention to detail and also to the overall structure of the art piece. Many different visual aspects can be useful in this task, for example the shape, the technique, the object type, and the motif, among others. Significantly, all these features are found at different levels of detail in images, which makes this a challenging problem, but also an appropriate one for HR images.

3.4 Super Resolution

To create the super resolution dataset (MetH-SR) we focus on the largest images available. We keep only those greater than 12 megapixels, which results in a total of 20,000 HR images. From these 20,000, 16,000 are used for training and 4,000 for test. Filtering by image size slightly reduces the variance in aspect ratios (see Figure 2). However, the dataset remains balanced between landscape and portrait images.

The super resolution task consist on generating a high-resolution image from its corresponding low-resolution version. In MetH-SR we propose super resolution task at four different scaling factors: x2, x4, x8 and x16. For the training subsets we do not include the Low Resolution (LR) images, only the original HR ones. We expect researchers to down-sample images for training their models as they consider best. In contrast, to guarantee a consistent performance evaluation, we provide the LR images for the test subset at all four scaling factors. For generating the LR version of images we perform a bicubic down-sampling. Notice that our most aggressive down-sampling, by a factor of x16, is unprecedented in the literature, but nevertheless results in images ranging between 46,284 and 80,316 pixels (except for the largest image which has 151,872 pixels). This range includes currently popular input sizes, like 224x224 and 256x256.

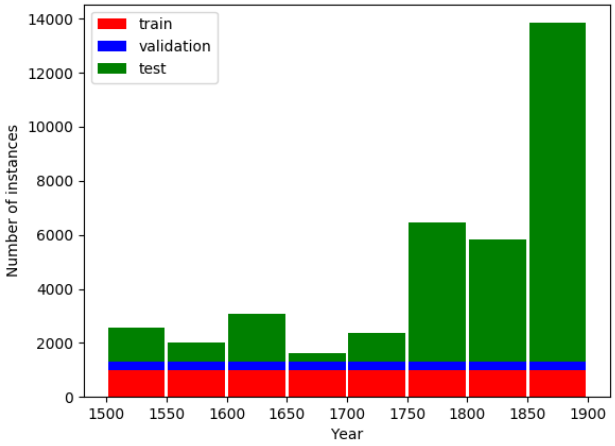

Figure 5. MetH-Period distribution of samples in bins of 50 years. Train, validation and test subsets are highlighted in red, blue and green respectively. Train and validation subsets are balanced within bins of 50 years. Test set contains the rest of instances and is left unbalanced.

4 BASELINES

In this section, we present a series of baselines for the introduced datasets, except for the MetH-SR. These should serve as a reference for researchers working on the proposed tasks. All baselines reported here down-sample the images from the MetH datasets, as it is not within the scope of this paper to provide contributions on working with high-resolution and variable-shape images. This would, and will, require a paper on its own. We expect solutions exploiting the full resolution of images to surpass these baselines by avoiding the loss of information and the deformation of the data.

The baselines of both image classification datasets (MetH-Medium and MetH-Cultures) and the regression dataset (MetH-Period), are rather similar. All of them use the Vgg16 architecture [47] randomly initialized. We decided to use Vgg16 over other architectures for its design simplicity, following the typical CNN structure (*i.e.*, a set of convolutional layers followed by a set of fully-connected ones). This architecture has achieved competitive results in challenging problems, for example, in Imagenet it reached 74.4% top-1 and 91.9% top-5 accuracy [47]. The only change in the architecture is in the output neurons, which have been adapted to each problem (the number of classes for the classification tasks, one for the regression task). Images are downsampled to 256x256, and a random crop of 224x224 is used during training. For validation and testing, instead of a random crop, a central crop is used. The baseline include the Dropout layers [48] used in the original Vgg16 architecture (applied after each fully-connected layer), with a rate of 0.5. For learning, Adam [27] is used with a learning rate of 0.001 on all three tasks.

The results obtained by all baselines are shown in Table 2. The MetH-Medium baseline obtained a 62.8% test accuracy after training the model for 24 epochs. The MetH-Cultures baseline obtained a 51.8% test accuracy after training the model for 39 epochs. The MetH-Period baseline obtained a 173.7 MSE in the test set after 36 epochs. All baseline models and the code needed to train them are

Dataset	Task	Splits (train/validation/test)	Target	Baseline Performance	Architecture
MetH-Medium	Classification	23,000 / 2,300 / 2,300	23 classes	Test accuracy: 62.8%	Vgg16
MetH-Cultures	Classification	12,000 / 1,500 / 1,500	15 classes	Test accuracy: 51.8%	Vgg16
MetH-Period	Regression	8,000 / 10,735 / 19,035	Range of years: 1500 to 1900	Test MSE: 173.7	Vgg16
MetH-SR	Super resolution	18,000 / - / 4,000	Scaling factors: x2, x4, x8, x16	-	-

Table 2. Baselines table for the MetH datasets. It includes general information for each task and its baseline results. MetH-SR has no benchmark.

publicly available [7].

5 DATASET IMPACT

Down-sampling images before training a CNN has become a frequent approach. This enforces a single shape across the dataset (removing restrictions from the model design), and reduces the computational cost of training. However, down-sampling always implies losing information, and in most cases implies deforming the remaining one. This work is based on the hypothesis that down-sampling will soon become the exception and HR the rule.

To promote research on solutions that make the most out of HR images with variable-shape, we introduce the MetH datasets. We strongly believe that there are solutions to be discovered which will make CNNs capable of taking advantage of these properties, obtaining greater performance than regular techniques which avoid HR and variable-shape properties simply by down-sampling data. To support this hypothesis, in this section we discuss some of the current challenges in the field of AI when working with HR and variable-shape properties, and some contributions that may already be useful in that regard.

Nevertheless, the challenge of using HR images for CNNs goes beyond the field of AI. Its technical requirements are so high, that state-of-the-art technology is currently a limiting factor. In order to properly tackle this challenge, the high-performance computing (HPC) field must contribute as well. HPC must produce advances that remove constrains from the set of solutions the field of AI may come up with. For this reason, we also analyze the challenges, contributions and future research lines that MetH datasets can help promote within the field of HPC.

5.1 Challenges for AI

Working with variable-shape is challenging at several levels. Traditional CNN architectures (*i.e.*, those composed by a set of convolutional and pooling layers followed by a set of fully-connected layers) can only be trained with a dataset of homogeneously shaped images. This is caused by the connection between the last convolutional layer and the first fully-connected layer. If this connection is not fixed, the number of parameters of the fully-connected becomes dependent on the output size of the convolutional, which at the same time depends on the size of the input. Thus, if the size of the input changed, the number of parameters in the fully-connected layer would also change along training. That is something unfeasible at the moment, but explorable through MetH. Notice how convolutional layers do not have this issue. These layers have a fixed number of parameters regardless of input size. Indeed, convolutional layers are not affected by variable-shape, but they do not fix it either; the output shape of convolutional layers changes with the input size.

A popular workaround for this issue is to add a Global Average Pooling (GAP) layer [31] between the convolutional and the fully-connected blocks. The GAP produces a fixed sized output, by reducing the output of each convolutional filter (*i.e.*, each channel of activations) to a single value. In this process, all spatial information is lost. This is not ideal, since spatial information may be of relevance for the final classifier. To avoid the complete loss of spatial information, the Spatial Pyramid Pooling (SPP) layer [21] represents as an intermediate approach. This layer fixes the output size but builds a representation that contains some relative spatial information. Approaches like GAP and SPP could be thoroughly evaluated and improved through the MetH challenges.

Even if we use architectures which can handle variable-shape images, we still require a consistent size for the images processed together in the same batch. Otherwise the output tensor would not be a regularly shaped one, as needed so far by convolutional neurons. In this context, to train with variable-shape images one must use a batch size of one, or online learning, which is sub-optimal in terms of learning and computational efficiency, and is an open field of research on its own [44]. A straight-forward solution is the use of padding to fill the gaps between the different image shapes in the same batch. However, if images are not similar in shape, a lot of padding may be needed. This could become a hindrance for learning, while also causing a significant computational overhead [18]. An alternative would be to use variable batch sizes, with the goal of constraining the amount of padding needed. This seems like a feasible approach in theory, as shown by recent contributions [19]. However, its performance in large-scale practice remains to be assessed, as well as the required shape-aware batching policy.

Another question is how to adapt the design for the learning needs of variable-shape datasets. Extreme shape variations are more demanding in terms of generalization (patterns may be very tall or very wide, even larger than the input), capacity (more and more efficient filters may be needed) and scale changes. Recent contributions could already be applied to mitigate some of these issues, like the SPP, Dilated filters [58] and attentional mechanisms [52]. Indeed, the capacity of attentional mechanisms of focusing on a small but relevant portion of the input seems particularly appropriate for HR.

5.2 Challenges for HPC

AI has become a compute demanding field, as a result of the recent increase in data availability and the popularization of DL. For many AI applications single CPUs are no longer able to satisfy the computational requirements of the field, leading to the convergence of HPC (*i.e.*, High-Performance Computing) and AI research. DL methods are currently deployed in parallel systems composed by multiple

[7] https://github.com/HPAI-BSC/MetH-baselines

nodes, where each node may contain a number of accelerators (*e.g.*, GPUS, TPUs, FPGAs) or multi-core CPUs [29, 34, 45, 46, 54]. However, most contributions so far do not directly tackle the main limitation when working with HR images: memory requirements.

Processing HR images generates huge internal tensor representations of activations, which can easily surpass the memory capacity found in current accelerators. At the moment, all possible solutions are sub-optimal. Either we work with dedicated accelerators with reduced memory capacity, or we use general purpose hardware with large memory spaces [54]. To provide a reference on the scale of the issue, consider that processing a single image (batch size of one) of 4,000x4,000 through the Vgg16 architecture in inference requires over 17.5GB of RAM. Processing the same image for training would require approximately 40GB of memory, assuming single precision (32-bit) floats.

One approach within HPC targets the reduction of memory requirements, instead of increasing memory capacity. This is typically implemented through half-precision formats for the mixed precision training of neural networks [36] (*i.e.*, bfloat16, fp16 and 16-bit integer based). Among those, bfloat16, which was conceived for DL training, has become the numeric alternative because of its wider dynamic range and smaller footprint [1, 10, 23, 25]. The benefit of using these operations is twofold. First, in terms of efficiency, a single instruction is capable of computing twice the operations when compared to the single precision counterpart, reducing the execution time. Second, the usage of half-precision operators reduces in half the required memory footprint addressing the memory capacity issue. In this context, MetH datasets provide a framework of experimentation where half-precision formats are not a toy technology to play with, but an indispensable necessity.

HPC is also moving towards satisfying larger memory requirements. One of the most traditional approaches involves paging techniques, coordinated by the hardware and the operating system. This involves the use of a secondary storage as an extension of main memory providing a larger address space. Unfortunately, secondary storage is orders of magnitude slower than memory (SSD latency ranges in the tens of microseconds while DRAM latency ranges in the hundreds of nanoseconds) and thus becomes impractical due to the performance degradation. Another option is to use 3D XPoint non-volatile memory, as proposed by Intel® through the Optane™DC Persistent Memory modules. These memory modules provides larger capacity than regular DRAM modules (up to 3 TiB per socket) at the cost of being only one order of magnitude slower than DRAM. This memory supports volatile usage, in which the DRAM acts as a last-level cache, and it lets applications use the huge address space without having to modify them. The performance of these emerging technologies in a stressful, large-scale setting like the one proposed in this work, remains to be assessed.

One more aspect in which HPC is challenged by MetH regards load balancing. The use of variable-shaped images inherently implies variable computational and memory requirements. To enable an efficient use of resources, one of the main concerns of HPC, current solutions need to be adaptable and flexible in real-time: consecutive batches of data may have very different requirements. If this issue is not addressed, particularly in a distributed computing environment, the slowest component will drive the runtime and limit scalability. In this context, we propose the use of variable-shape images, as provided by MetH, as a prioritary case study for load balacing research.

5.3 Art is Rich

Art is a tricky domain to work with. It is creative, but also strongly influenced by context and previous contributions. There are some general aesthetic rules, however these are often broken as part of the creative process. As we will review in this section, working with art datasets provides yet another level of relevance to the MetH datasets.

Art is scale independent. Let us consider the case of the fleur-de-lis, a recurrent motif in French art. This is a useful thing to know when assessing the cultural origin of an art piece, or even the creation date (the fleur-de-lis was a symbol of royal France). In art, the fleur-de-lis can be found in a variety of ways. It can be shown large on tapestries and canvases, or it can be a tiny decoration on silverware and coins. Such scale variance represents a challenge for DL, and so far it is hard to find other domains where complex patterns can have such a wide variance in scale.

Art is hard to predict. All MetH datasets have a large intra-class variability. Indeed, objects made from the same material have significant differences among them. As do objects belonging to the same culture or objects coming from the same period. At the very least, the MetH-Medium includes cultural and temporal variations. The MetH-Cultures, medium and temporal variations. And the MetH-Period, medium and cultural variations. Super resolution tasks in the context of art is also affected by art unpredictability. Producing credible extrapolations to produce HR images from LR images is particularly challenging on a domain based on creativity.

Art can be blended. To solve all the MetH tasks optimally, one needs to find the common trends found within the same materials, cultures and periods. If these shared properties are learnt, they can be enforced on other data through style transfer [16]. This methods will allow us to produce interesting visual fusions. What if Picasso had been born in China? What if the Mona Lisa was sculpted of stone? Such research could result in applications for art dissemination, usable by museums world-wide.

Art is pervasive. It stands to reason that, among all the visual features needed to solve all the MetH tasks, there will be a significant coherency. In this scenario, the MetH datasets become an interesting use case study for lifelong learning techniques [40]. These methods aim at training models capable of solving several tasks, and seem particularly appropriate for a set of tasks with very different goals, but very similar data. Following the same assumption, the transfer learning field [15, 57] will also find an interesting testing ground in the MetH datasets. Transfer learning aims at improving the performance on one task by reusing knowledge obtained when trying to solve a different one. In this context, one could use features learnt for MetH-Cultures to solve MetH-Medium, or any other combination.

6 CONCLUSIONS

Currently, the default practice for training CNNs has been to systematically down-sample data, homogenizing samples. This simplifies architecture design and reduces computational cost, but implies information loss and deformation. For HR data this can be dramatic, as down-sampling a 4,000x4,000 image to 256x256 pixels implies losing 99.6% of all data. This methodology has been successful when tackling the first generation of DL tasks, but it will be inefficient or even unfeasible for more ambitious future tasks, which may require both attention to detail and an understanding of the overall structure.

The goal of this paper is to encourage people to work with datasets of HR and variable-shape, exploiting all the information available and avoiding deformation. We expect this approach will eventually

lead to better performances. To promote this line of research we introduce four novel datasets: MetH-Medium, MetH-Cultures, MetH-Period and MetH-SR. The first two are image classification problems, the MetH-Period is a regression problem and, the MetH-SR is a super resolution task. The four datasets contain images of art pieces from the Metropolitan Museum of Art of New York [39].

In comparison, public datasets available today are composed by significantly smaller images: In a sample of 12 popular datasets we found the largest image to have 1 MP (see left plot of Figure 2). In contrast, the smallest images in our datasets have 2.6 MP, while the average is 5 MP. Another difference between currently popular datasets and our proposal is the distribution of aspect ratios. The analyzed sample had a majority of landscape images, with only a few having a significant amount of portrait ones. As shown in the right plot of Figure 2, even in these few cases there is still a clear majority of landscape images. Our datasets have more balanced aspect ratio distributions, perfectly balanced in two cases and slightly biased towards portrait aspect ratios for the remaining two. These two characteristics (larger resolution and aspect ratio variance) are relevant enough as to have a significant impact on the performance of current CNNs methods. Thus, we argue that past and future contributions to the field should also be evaluated on this setting.

For all the proposed datasets except MetH-SR, we provide a baseline based on well-known architectures in the field. Although these are not thoroughly optimized, their performance is still rather mediocre as shown in Table 2. This shows that there is plenty of room for improvement, and that the problem is challenging: MetH-Medium contains a subset of classes particularly hard to discriminate, MetH-Cultures can only be discriminated through a wide variety of visual features and MetH-Period spans through 500 years and a lot of artistic movements. MetH-SR task reaches a degree of upsampling unprecedented in the literature, stating the challenge of the problem.

The proposed datasets are also relevant for a variety of reasons. Within the field of AI it can contribute to topics like input pipelines, new neuron and model designs, scaling and scalable architectures, online and batch learning. Within the field of HPC it can contribute to topics like accelerator design, memory technologies, paging techniques and load-balancing. Considering the particularities of art, the dataset provided can also be used for research related with style transfer, lifelong learning and transfer learning.

All four datasets, the instructions to use them, and the code and results for the baselines are made available through a publicly available webpage.

ACKNOWLEDGEMENTS

This work is partially supported by the Intel-BSC Exascale Lab agreement, by the Spanish Government through Programa Severo Ochoa (SEV-2015-0493), by the Spanish Ministry of Science and Technology through TIN2015-65316-P project, and by the Generalitat de Catalunya (contracts 2017-SGR-1414).

REFERENCES

[1] 'A Dynamic Approach to Accelerate Deep Learning Training'. Submitted to International Conference on Learning Representations, 2020.

[2] Beyond imagenet large scale visual recognition challenge. `http://image-net.org/challenges/beyond_ilsvrc`. Accessed: 2019-11-14.

[3] Contributing data to deepfake detection research. `https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html`. Accessed: 2019-11-12.

[4] Creating a data set and a challenge for deepfakes. `https://ai.facebook.com/blog/deepfake-detection-challenge/`. Accessed: 2019-11-12.

[5] Deepfake detection challenge (dfdc). `https://deepfakedetectionchallenge.ai/`. Accessed: 2019-11-12.

[6] Met museum image and data resources. `https://www.metmuseum.org/about-the-met/policies-and-documents/image-resources`. Accessed: 2019-11-14.

[7] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik, 'Contour detection and hierarchical image segmentation', *IEEE transactions on pattern analysis and machine intelligence*, **33**(5), 898–916, (2010).

[8] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel, 'Low-complexity single-image super-resolution based on nonnegative neighbor embedding', (2012).

[9] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool, 'Food-101–mining discriminative components with random forests', in *European Conference on Computer Vision*, pp. 446–461. Springer, (2014).

[10] N. Burgess, J. Milanovic, N. Stephens, K. Monachopoulos, and D. Mansell, 'Bfloat16 Processing for Neural Networks', in *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, pp. 88–91, (June 2019).

[11] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia, 'Multi-view 3d object detection network for autonomous driving', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, (2017).

[12] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi, 'Describing textures in the wild', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, (2014).

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[14] Common Visual Data Foundation. Google landmarks v2 dataset. `https://github.com/cvdfoundation/google-landmark#release-history`, 2019.

[15] Dario Garcia-Gasulla, Armand Vilalta, Ferran Parés, Eduard Ayguadé, Jesus Labarta, Ulises Cortés, and Toyotaro Suzumura, 'An out-of-the-box full-network embedding for convolutional neural networks', in *2018 IEEE International Conference on Big Knowledge (ICBK)*, pp. 168–175. IEEE, (2018).

[16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, 'Image style transfer using convolutional neural networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, (2016).

[17] Krzysztof J Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, and Kyunghyun Cho, 'High-resolution breast cancer screening with multi-view deep convolutional neural networks', *arXiv preprint arXiv:1703.07047*, (2017).

[18] Swarnendu Ghosh, Nibaran Das, and Mita Nasipuri, 'Reshaping inputs for convolutional neural network: Some common and uncommon methods', *Pattern Recognition*, **93**, 79–94, (2019).

[19] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He, 'Accurate, large minibatch sgd: Training imagenet in 1 hour', *arXiv preprint arXiv:1706.02677*, (2017).

[20] Gregory Griffin, Alex Holub, and Pietro Perona, 'Caltech-256 object category dataset', (2007).

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Spatial pyramid pooling in deep convolutional networks for visual recognition', *IEEE transactions on pattern analysis and machine intelligence*, **37**(9), 1904–1916, (2015).

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).

[23] Greg Henry, Ping Tak Peter Tang, and Alexander Heinecke. Leveraging the bfloat16 Artificial Intelligence Datatype For Higher-Precision Computations, 2019.

[24] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, 'Single image

super-resolution from transformed self-exemplars', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, (2015).

[25] Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, Jiyan Yang, Jongsoo Park, Alexander Heinecke, Evangelos Georganas, Sudarshan Srinivasan, Abhisek Kundu, Misha Smelyanskiy, Bharat Kaul, and Pradeep Dubey. A study of BFLOAT16 for Deep Learning Training, 2019.

[26] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li, 'Novel dataset for fine-grained image categorization: Stanford dogs', in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, (2011).

[27] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).

[28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, 'Hmdb: a large video database for human motion recognition', in *2011 International Conference on Computer Vision*, pp. 2556–2563. IEEE, (2011).

[29] Thorsten Kurth, Sean Treichler, Joshua Romero, Mayur Mudigonda, Nathan Luehr, Everett Phillips, Ankur Mahesh, Michael Matheson, Jack Deslippe, Massimiliano Fatica, Prabhat, and Michael Houston, 'Exascale Deep Learning for Climate Analytics', in *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, SC '18, pp. 51:1–51:12, Piscataway, NJ, USA, (2018). IEEE Press.

[30] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari, 'The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale', *arXiv:1811.00982*, (2018).

[31] Min Lin, Qiang Chen, and Shuicheng Yan, 'Network in network', *arXiv preprint arXiv:1312.4400*, (2013).

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, 'Microsoft coco: Common objects in context', in *European conference on computer vision*, pp. 740–755. Springer, (2014).

[33] William Lotter, Greg Sorensen, and David Cox, 'A multi-scale cnn and curriculum learning strategy for mammogram classification', in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 169–177, Springer, (2017).

[34] Amrita Mathuriya, Thorsten Kurth, Vivek Rane, Mustafa Mustafa, Lei Shao, Debbie Bard, Prabhat, and Victor W. Lee, 'Scaling GRPC TensorFlow on 512 nodes of Cori Supercomputer', *CoRR*, **abs/1712.09388**, (2017).

[35] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa, 'Sketch-based manga retrieval using manga109 dataset', *Multimedia Tools and Applications*, **76**(20), 21811–21838, (2017).

[36] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al., 'Mixed precision training', *arXiv preprint arXiv:1710.03740*, (2017).

[37] Vinod Nair and Geoffrey E Hinton, 'Rectified linear units improve restricted boltzmann machines', in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, (2010).

[38] Maria-Elena Nilsback and Andrew Zisserman, 'Automated flower classification over a large number of classes', in *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, pp. 722–729. IEEE, (2008).

[39] The Museum of Modern Art. Open access initiative github repository. `https://github.com/metmuseum/openaccess`, c2db720f0ecd33db7bdf920fd032b62f1c50626a, 2016.

[40] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter, 'Continual lifelong learning with neural networks: A review', *Neural Networks*, (2019).

[41] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar, 'Cats and dogs', in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3498–3505. IEEE, (2012).

[42] Ariadna Quattoni and Antonio Torralba, 'Recognizing indoor scenes', in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 413–420. IEEE, (2009).

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., 'Imagenet large scale visual recognition challenge', *International journal of computer vision*, **115**(3), 211–252, (2015).

[44] Doyen Sahoo, Quang Pham, Jing Lu, and Steven CH Hoi, 'Online deep learning: Learning deep neural networks on the fly', *arXiv preprint arXiv:1711.03705*, (2017).

[45] Alexander Sergeev and Mike Del Balso, 'Horovod: fast and easy distributed deep learning in TensorFlow', *CoRR*, **abs/1802.05799**, (2018).

[46] S. Shi, Q. Wang, P. Xu, and X. Chu, 'Benchmarking State-of-the-Art Deep Learning Software Tools', in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pp. 99–104, (Nov 2016).

[47] Karen Simonyan and Andrew Zisserman, 'Very deep convolutional networks for large-scale image recognition', *arXiv preprint arXiv:1409.1556*, (2014).

[48] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, 'Dropout: a simple way to prevent neural networks from overfitting', *The journal of machine learning research*, **15**(1), 1929–1958, (2014).

[49] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, 'Going deeper with convolutions', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, (2015).

[50] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang, 'Ntire 2017 challenge on single image super-resolution: Methods and results', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 114–125, (2017).

[51] Michael Treml, José Arjona-Medina, Thomas Unterthiner, Rupesh Durgesh, Felix Friedmann, Peter Schuberth, Andreas Mayr, Martin Heusel, Markus Hofmarcher, Michael Widrich, et al., 'Speeding up semantic segmentation for autonomous driving', in *MLITS, NIPS Workshop*, volume 2, p. 7, (2016).

[52] 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*, pp. 5998–6008, (2017).

[53] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera, 'Rgb-d-based human motion recognition with deep learning: A survey', *Computer Vision and Image Understanding*, **171**, 118–139, (2018).

[54] Gu-Yeon Wei, David Brooks, et al., 'Benchmarking TPU, GPU, and CPU Platforms for Deep Learning', *arXiv preprint arXiv:1907.10701*, (2019).

[55] Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang, 'Ip102: A large-scale benchmark dataset for insect pest recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8787–8796, (2019).

[56] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2019.

[57] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, 'How transferable are features in deep neural networks?', in *Advances in neural information processing systems*, pp. 3320–3328, (2014).

[58] Fisher Yu and Vladlen Koltun, 'Multi-scale context aggregation by dilated convolutions', *arXiv preprint arXiv:1511.07122*, (2015).

[59] Roman Zeyde, Michael Elad, and Matan Protter, 'On single image scale-up using sparse-representations', in *International conference on curves and surfaces*, pp. 711–730. Springer, (2010).

[60] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, 'Places: A 10 million image database for scene recognition', *IEEE transactions on pattern analysis and machine intelligence*, **40**(6), 1452–1464, (2017).