Preprint submitted to Acta Materialia on 09 January 2017

Exploring the microstructure manifold: image texture representations applied to ultrahigh carbon steel microstructures

Brian L. DeCost, Toby Francis, Elizabeth A. Holm

February 10, 2017

Abstract

We introduce a microstructure informatics dataset focusing on complex, hierarchical structures found in a single Ultrahigh carbon steel under a range of heat treatments. Applying image representations from contemporary computer vision research to these microstructures, we discuss how both supervised and unsupervised machine learning techniques can be used to yield insight into microstructural trends and their relationship to processing conditions. We evaluate and compare keypoint-based and convolutional neural network representations by classifying microstructures according to their primary microconstituent, and by classifying a subset of the microstructures according to the annealing conditions that generated them. Using t-SNE, a nonlinear dimensionality reduction and visualization technique, we demonstrate graphical methods of exploring microstructure and processing datasets, and for understanding and interpreting high-dimensional microstructure representations.

1 Introduction

Comprised of the structures that arise from processing and mediate properties, microstructure is a core focus of the discipline of materials science. Microstructural information is most often conveyed via images obtained through various microscopy techniques (i.e. micrographs), sometimes supplemented by other structural and compositional probes. Traditionally, microstructural images have been evaluated by human experts, both to interpret the micrographs themselves and to connect them to processing conditions and property outcomes. However, recent research in microstructure informatics

has begun to explore applications of contemporary computer vision to construct microstructure representations suitable for use in machine learning and microstructure analytics tasks[1, 2, 3, 4]. For example, [3] compare several image texture representations and find that off-the-shelf convolutional neural network (CNN) features can be applied to microstructure analytics tasks (e.g. classification) without fine-tuning any of the CNN parameters. Likewise, Lubbers et al.[4] apply bilinear CNN representations[5, 6, 7] to synthetic lamellar structures, and relate this representation to the generative microstructure model parameters (i.e. lamellar spacing and orientation and noise). While promising proofs of principle, these studies used comparatively simple and well-parameterized microstructures. To move towards quantitative application of generic computer vision techniques, we require real-world, technologically-relevant microstructure systems exhibiting the complex, hierarchical structures that challenge conventional microstructure segmentation and quantification.

To this end, we introduce the CMU-UHCS (Carnegie Mellon University Ultrahigh Carbon Steel) dataset[1], based on the work of Hecht et al.[8, 9]. This dataset consists of 961 scanning electron microscopy (SEM) micrographs of Ultrahigh Carbon Steel (UHCS) subjected to a variety of heat treatments and taken at several different magnifications. The dataset spans several complex and hierarchical microconstituents typically found in UHCS and other technologically relevant alloy systems, offering a compelling real-world microstructure informatics challenge.

UHCS (steels with 1-2.1 wt% carbon) are intermediate in content to high carbon steel (0.6-1 wt% C) and cast iron (2.1-4.3 wt% C). Due to their high carbon content relative to conventional steels, a characteristic microstructure feature of these alloys is proeutectoid cementite (Fe_3C), typically forming a carbide network associated with the grain boundaries of the high-temperature austenite phase. The hard, brittle carbides help lend UHCS its well-known high strength and wear resistance, but highly-connected intergranular carbide networks can be detrimental to toughness and ductility by providing extended pathways for crack propagation[10, 11]. Recent UHCS research has focused on mitigating this weakness by optimizing the network microstructure through various heat treatments[12] and addition of minor alloying elements[13, 14]. Hecht et al. recently developed a quantitative measure of the carbide network connectivity, relating this to annealing schedules and toughness measurements[8]. A similar study concerning the effect of annealing conditions on spheroidite morphology is forthcoming[9]. The present UHCS

[1] to appear on `https://materialsdata.nist.gov`

microstructure dataset is built on the characterization efforts for these two UHCS studies.

In this study, we use the UHCS dataset compare state-of-the-art CNN-based image texture representations with the classic bag of visual words (BoW) representation[15, 16]. As microstructure representations, the BoW has the theoretical advantage of strong explicit scale and rotation invariance, while CNNs notoriously outperform BoW on typical natural image recognition tasks (e.g. facial recognition, object detection and identification, scene classification). We evaluate each image representation using both supervised and unsupervised learning methods, and demonstrate how these techniques can be used together for exploratory microstructure analysis. Specifically, we used a Support Vector Machine (SVM)[17] approach to classify microstructures both by primary microconstituent and annealing condition. We complement this understanding by applying the unsupervised dimensionality-reduction technique t-SNE (t-distributed Stochastic Neighbor Embedding)[18] to visualize the high-dimensional distributions of each microstructure representation, relating this structure to available annealing schedule and imaging metadata.

Our primary contributions in this report are:

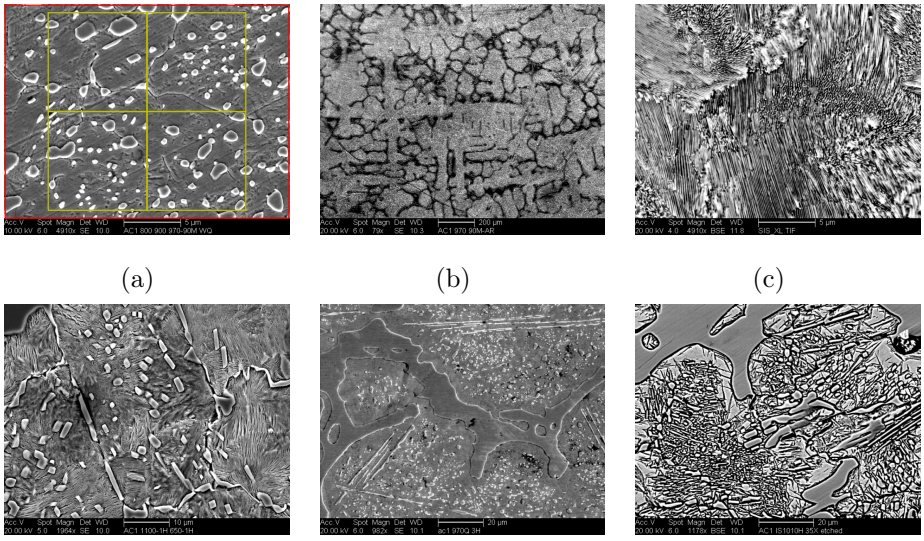
- A real-world dataset of complex, hierarchical microstructures annotated with microstructure constituent metadata, as well as partial imaging and processing metadata, such as heat treatment, quenching procedure, and magnification.
- Evaluation of several competitive computer vision techniques, with discussion of their relative strengths and weaknesses for a range of real-world microstructure informatics tasks.
- Exploration of these microstructure representations for inferring processing – microstructure – properties relationships for realistic complex, hierarchical microstructure systems.

2 Methods

2.1 UHCS Dataset

The UHCS dataset consists of 961 SEM micrographs of commercial UHCS subjected to a range of heat treatments by Hecht et al.[8, 9]. These micrographs span a wide range of magnifications, and include both secondary electron (SE) and back-scattered electron (BSE) images. 598 micrographs also have annealing schedule metadata: annealing time, temperature, and

quench medium. All 961 images are labeled with their primary microstructure constituents as illustrated in Figure 1. Most of the micrographs focus on the spheroidite morphology (Figure 1a), the carbide network (Figure 1b), and pearlite (Figure 1c). A smaller number of micrographs contain two primary microconstituents, such as pearlite containing spheroidite (Figure 1d), Widmanstätten cementite (Figure 1e), and martensite (Figure 1f). Table 1 shows the distribution of each of these primary microconstituent labels. We used the full set of 961 labeled 645×484 pixel micrographs to generate the data visualizations in Section 3.2.



(a) (b) (c)

(d) (e) (f)

Figure 1: Primary microstructure constituents in the UHCS dataset: (a) spheroidized cementite with red and yellow frames indicating image regions used for feature extraction in the UHCS-600 and UHCS-2400 datasets, (b) carbide network microstructure, (c) pearlite, (d) pearlite containing spheroidized cementite, (e) Widmanstätten cementite, and (f) martensite and/or bainite.

For the primary microconstituent classification experiments, we considered only a subset of these labeled micrographs: 200 randomly selected micrographs each from the spheroidized cementite, carbide network, and pearlite/pearlite+spheroidite classes, for a total of 600 images. We also consider an expanded dataset constructed by cropping four 224×224 subimages from the center of each micrograph in the original dataset, so that the expanded dataset consists of 2400 images. The single red and four yellow

Table 1: Schedule of primary microconstituent labels in the UHCS micrograph dataset.

primary microconstituents	# of micrographs
spheroidite	374
carbide network	212
pearlite	124
pearlite + spheroidite	107
Widmanstätten cementite	81
pearlite + Widmanstätten	27
Martensite/Bainite	36

frames in Figure 1a indicate the image regions used for microstructure feature extraction in the full-sized and cropped image sets, respectively.

The annealing schedule classification task was limited to the micrographs collected to study the spheroidite morphology. The dataset contains spheroidite micrographs resulting from 23 distinct annealing schedules. Within this subset of micrographs, we limit the classification dataset to the 13 annealing conditions with at least 15 micrographs. Where more than 15 micrographs with a given annealing condition are available, we randomly select 15 micrographs to obtain a balanced classification dataset. The resulting annealing condition classification datasets consist of 195 full-sized micrographs and 780 cropped micrographs.

2.2 Image representations

In this work, we explore and compare two computer vision approaches for computing generic image representations: Mid-level image patch descriptors[16, 19] and convolutional neural network (CNN) representations[20, 21, 22]. The mid-level features approach is attractive due to its relatively strong invariance to image scale and orientation; its focus on identifying and characterizing individual features is also intuitive to the materials scientist. However, CNN representations are generally regarded as richer, more hierarchical, and more effective than mid-level image features, even when transferring CNN parameterizations from one task to another (in this case completely unrelated) task.

2.2.1 Mid-level image features

The baseline feature extraction method in this study is the bag of visual words (BoW) method, which represents an image as a distribution of local image descriptors (i.e. visual features). As previously reported in detail[1, 2], we applied both the Difference of Gaussians[23] and the Harris-LaPlace[24] interest point detectors to select distinctive image regions with characteristic scales and orientations. We then used oriented SIFT descriptors[23] to characterize the visual appearance of each interest point, and k-means clustering[25] to quantize the SIFT descriptors into a visual dictionary with 100 (BoW_{100}) visual words (i.e. SIFT cluster centers). Each image is then represented by its microstructural fingerprint: A normalized histogram measuring the occurrence frequency of each visual word within the image. BoW representations can be compared with various similarity metrics for discrete probability distributions, such as the Hellinger and χ^2 kernels[16]. Presently, we use the χ^2 kernel, which for two normalized m-dimensional histograms $X = \{x_i\}$ and $Y = \{y_i\}$ is written as:

$$D_{\chi^2}(X, Y) = \frac{1}{2} \sum_{i=1}^{m} \frac{(x_i - y_i)^2}{x_i + y_i} \tag{1}$$

Typically the exponential version of the χ^2 is used with SVM classification:

$$K_{\chi^2}(X, Y) = \exp\left(-\frac{1}{A} D_{\chi^2}(X, Y)\right) \tag{2}$$

We follow [16] by setting the kernel parameter A to the mean χ^2 distance between BoW representations of all of the training examples.

In addition to sparse, oriented SIFT features computed at interest points (s-SIFT), we also used dense sampling of SIFT features (d-SIFT) at multiple scales, with fixed orientation. This results in a much larger set of local features for constructing BoW representations and lends itself to more efficient numerical implementation[26]. In some applications, this can lead to improved recognition performance, even though the individual region descriptors are no longer rotation invariant.

Ambiguity between visual words is a major weakness of the BoW methods: in general, boundaries between clusters in the visual dictionary are arbitrary[27] in the sense that they are a convenient means of discretizing a high-dimensional vector space rather than based on a physically meaningful set of visual features[28]. The result is that a feature located near a cluster boundary may be very far from representative of that cluster of features, yet

it is weighted the same as a feature near the cluster centroid. Worse still, a very similar feature that falls on the other side of the cluster boundary will be assigned to a different visual word! There are many techniques that attempt to mitigate the resulting visual word assignment uncertainty, for example by assigning local feature descriptors to multiple visual words[29] or by using a probabilistic visual dictionary[27, 30]. Other methods record more detailed information about the relationships between the visual dictionary and local feature descriptors extracted from an image[31]. In this report, we explore one simple and effective method: VLAD (Vector of Locally Aggregated Descriptors)[32].

2.2.2 VLAD encoding

The Vector of Locally Aggregated Descriptors (VLAD)[32] technique is closely related to the BoW method. VLAD attempts to mitigate the ambiguity between visual words by recording the *difference* between local feature descriptors and the corresponding visual word (i.e. the cluster centroid), rather than simply constructing a distribution of visual word occurrence frequency. After assigning all the local feature descriptors to visual words, VLAD sums up the residual vectors (the differences) between each visual word and all the local features assigned to it. The final VLAD descriptor is obtained by concatenating the residual sums for each visual word. We applied the block-wise normalization scheme (intra-normalization) from [33]: each residual sum is L2-normalized before being concatenated, and the resulting VLAD vector is L2-normalized as well.

We used VLAD to encode both sparse and dense SIFT features extracted using the same methods outlined in Section 2.2.1. With VLAD encoding, it is common to use a smaller visual dictionary size; here we used dictionary sizes of 32 (VLAD_{32}) and 64 (VLAD_{64}). This results in VLAD features with $128 \times 32 = 4096$ and $128 \times 64 = 8192$ dimensions, respectively.

2.2.3 Convolutional Neural Network features

In the past few years, convolutional neural networks (CNN) have demonstrated excellent performance at many computer vision tasks; this success is often credited to the hierarchical nature of the image representation they construct. CNNs extract high-level image features by stacking multiple layers of neurons organized into convolution filters learned from annotated training images. By interleaving pooling (effectively down-sampling) steps between layers of convolution filters, CNNs obtain hierarchical representations of image content.

Though CNNs are notorious for requiring extreme amounts of training data (and for overfitting on small datasets), recent research efforts in *transfer learning*[34] have shown that deep CNNs can generalize well to new datasets, in some cases even when new task is not related to the original task[35]. Simple approaches include using the output of the high-level CNN layers as input to a linear SVM[36], or retraining (fine-tuning) some or all of the layers of the pre-trained CNN using a new training set. In this study we use high-level features the from the VGG16 CNN architecture[37], parameterized for object recognition on the ImageNet ILSVRC-2014 dataset[38], which consists of approximately 1.2 million images representing 1000 object categories (none of which include microstructures). The VGG16 architecture consists of 14 convolution layers arranged into 5 blocks delineated by pooling (upsampling) layers, followed by two fully-connected layers of 4096 neurons each, and a final 1000-class classification layer. Because the VGG16 CNN operates on color images, we preprocess each SEM micrograph by replicating the raw grayscale image in each color channel of a new RGB image and subtracting the average intensity of the ImageNet training set for each channel, as recommended by [37]. We used the publicly available parameters provided by the VGG group[37] without any fine-tuning.

Fully-connected CNN features were not computed for the full-sized images, because the fully-connected layers fix the size of allowable input images to the size of the training set images. This can be mitigated by pooling the fully-connected CNN features from appropriately-sized regions within a larger image[39]. However, for transfer learning tasks (and particularly for image texture recognition) it is much more efficient and effective to apply pooling to the high-level convolution layers[40], which can be easily extracted from input images of arbitrary size.

We investigated the third convolution layer from both the fourth (VGG_4) and fifth (VGG_5) convolution blocks of the VGG16 architecture. For this neural network, both the VGG_4 and VGG_5 convolution blocks produce 512-channel feature maps, respectively sized 14×14 and 7×7 for the cropped UHCS input images and 40×30 and 20×15 for the large UHCS input images. We used VLAD encoding with both a 32-element ($VLAD_{32}$) and a 64-element ($VLAD_{64}$) dictionary on the VGG_4 and VGG_5 feature maps, yielding VLAD vectors of length $512 \times 32 = 16384$ and $512 \times 64 = 32768$ respectively. One advantage of using an encoding method such as VLAD on convolution features is that it yields a deep representation of the image structure with no explicit high-level spatial dependence – a desirable property for image texture (and microstructure) recognition[40].

Finally, we explore a simple technique to increase the scale invariance of

these CNN representations. For each input image, we apply VLAD encoding to VGG_4 and VGG_5 feature maps from four scales, yielding multiscale CNN representations ($mVGG_4$ and $mVGG_5$). We use bilinear interpolation to downsample the original resolution twice by a factor of $\sqrt{2}$, and to upsample the original resolution once by the same factor. A greater degree of scale-invariance can be achieved by pooling over a finer-grained set of image scales.

2.3 SVM classification

We use Support Vector Machine (SVM) classification for microstructure categorization on the UHCS dataset. We compared BoW representations using the χ^2 kernel as outlined in Section 2.2.1. For each of the other image representations, we used linear SVM classification. Because SVM classification is sensitive to the absolute scale of the input features, we L2-normalize them and set the SVM margin parameter C to 1, following[40]. Reported performance figures are obtained via 10× 10-fold cross-validation on the full dataset; the uncertainties reported are sample standard deviations computed on the 100 validation sets. The classification results are insensitive to changes in the value of the margin parameter C.

2.4 Data visualization

We visualize each high-dimensional microstructure representation using t-SNE (t-distributed Stochastic Neighbor Embedding)[18], a non-parametric visualization technique for high-dimensional data. t-SNE often better captures high-dimensional structure of real-world data compared with other dimensionality reduction techniques such as principal component analysis (PCA)[41], multidimensional scaling (MDS)[42], Isomap[43], and Locally Linear Embedding[44]. Rather than preserving global distances between dissimilar points as in PCA, t-SNE preserves only the local structure and similarity of the data points, using a probabilistic measure of 'similarity'. t-SNE uses the high-dimensional Euclidean distance between data points to define pairwise conditional probabilities for each point being a close neighbor to all others in the dataset. The similarity of a high-dimensional data point $\mathrm{x_j}$ with respect to the data point $\mathrm{x_i}$ is modeled as a probability of observing $\mathrm{x_j}$ under a gaussian distribution P_i with variance σ_i centered at $\mathrm{x_i}$: $p_{j|i} \sim \exp\left(-||x_i - x_j||^2/2\sigma_i\right)$. The σ_i are chosen so that each gaussian distribution has a fixed perplexity $Perp(P_i) = 2^{H(P_i)} = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}}$, effectively tuning the number of nearest neighbor data points. The similarity of a map point $\mathrm{y_i}$ with respect to another map point $\mathrm{y_j}$ is analogous, replacing

the gaussian distribution with the student t distribution Q_i with one degree of freedom: $q_{j|i} \sim \left(1 + ||y_i - y_j||^2\right)^{-1}$.

t-SNE proceeds by attempting to minimize the mismatch in these conditional neighbor probabilities between the original high-dimensional dataset and the low-dimensional representation. This is accomplished by minimizing the sum of Kullback-Leibler divergences for each data point: $\min \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$ via gradient descent. This objective function emphasizes local structure by heavily penalizing large distances between map points where the corresponding high-dimensional distance is small, while effectively ignoring small distances between map points where the high-dimensional distance is large. Because t-SNE is a stochastic algorithm (unlike PCA), we perform t-SNE ten times for each image representation and select the best map, i.e. the map with the smallest value for the objective function. We use the same t-SNE map for each representation when drawing maps of class labels and processing parameters.

3 Results and Discussion

3.1 Primary microconstituent classification

In order to evaluate various microstructural representations, we classified the UHCS micrographs both by microconstituent categories and by annealing conditions. The microconstutuent classification was performed separately on the UHCS-600 dataset (600 full-size images, 200 in each of three categories) and the UHCS-2400 dataset (2400 cropped images, 800 in each of three categories). The annealing schedule classification examined the same datasets, but was limited to the spheroidite images produced by the 13 distinct annealing schedules with at least 15 micrographs. With larger image sets and fewer classes, the microconstutuent classification is more representative of the attainable accuracy of the methods, while the annealing condition classification challenges the approach in the limit of small datasets.

Table 2 reports cross-validation accuracies obtained via SVM classification using each of the feature extraction methods outlined in Section 2.2. The first two data columns show the average validation set accuracies on the UHCS-2400 and UHCS-600 image sets for the primary microconstituent classification task; the second two data columns show the same for the spheroidite annealing condition classification task. For the microconstituent classification task, accuracies are computed via 10× 10-fold cross-validation, while for the annealing condition classification task accuracies are obtained via a stratified leave-one-out cross-validation scheme. The uncertainties reported

are the standard deviations of the accuracies achieved on the validation and training sets, respectively. Uncertainties for the annealing schedule classification task are much higher, principally because of the much smaller dataset size.

The classification accuracy of a given feature extraction method is generally consistent between the two UHCS datasets, with a slight (within measurement uncertainty) advantage for the uncropped dataset. This advantage could potentially be a result of higher variability in the UHCS-2400 dataset, as not all of the original micrographs exhibit homogeneous microstructure features. For example, the cropped spheroidite images in Figure 1a are clearly not representative microstructure samples relative to the full image: the upper left quadrant contains a high proportion of grain boundary cementite, while the upper right quadrant contains almost no grain boundary cementite and a high proportion of interior cementite.

Table 2: Quantitative evaluation of microstructure representations. Cross-validation accuracy (± standard deviation) for SVM classification. The left two data columns show validation set scores for the primary microconstituent classification task, and the right two columns show the validation set scores for the annealing schedule classification task.

	microconstituent		annealing schedule	
method	UHCS-2400	UHCS-600	UHCS-2400	UHCS-600
raw	45.0 (± 5.36)	55.3 (± 5.73)	14.5 (± 3.07)	18.5 (± 9.99)
dSIFT BoW_{32}	86.9 (± 4.65)	89.0 (± 3.82)	48.7 (± 6.53)	39.5 (± 11.2)
dSIFT BoW_{100}	91.3 (± 3.76)	92.8 (± 2.87)	61.0 (± 7.32)	50.8 (± 16.1)
sSIFT BoW_{32}	88.8 (± 3.53)	91.2 (± 3.26)	59.9 (± 6.2)	41.0 (± 6.92)
sSIFT BoW_{100}	92.4 (± 3.36)	92.2 (± 3.39)	66.2 (± 5.67)	50.3 (± 9.13)
dSIFT $VLAD_{32}$	92.5 (± 3.01)	94.9 (± 2.85)	74.6 (± 5.25)	74.4 (± 9.04)
dSIFT $VLAD_{100}$	94.0 (± 2.55)	95.9 (± 2.59)	81.2 (± 8.0)	76.9 (± 11.3)
sSIFT $VLAD_{32}$	95.3 (± 2.46)	96.1 (± 2.31)	80.1 (± 4.03)	75.4 (± 12.1)
sSIFT $VLAD_{100}$	95.7 (± 2.31)	96.8 (± 2.42)	84.1 (± 5.36)	79.0 (± 8.95)
VGG_4 $VLAD_{64}$	–	97.9 (± 1.88)	–	84.6 (± 8.72)
VGG_4 $VLAD_{32}$	96.6 (± 2.31)	98.1 (± 1.78)	88.7 (± 3.32)	84.1 (± 6.14)
VGG_5 $VLAD_{32}$	94.7 (± 2.82)	98.5 (± 1.46)	76.4 (± 5.41)	78.5 (± 8.82)
VGG_5 $VLAD_{64}$	–	98.9 (± 1.17)	–	83.1 (± 6.63)
$mVGG_4$ $VLAD_{32}$	–	98.2 (± 1.71)	–	81.0 (± 9.58)
$mVGG_5$ $VLAD_{32}$	–	98.3 (± 1.57)	–	80.5 (± 8.66)

Using raw (normalized) flattened images as feature vectors affords around

50% accuracy on the microconstituent classification task – substantially better than the expected performance of a random classifier (33%). The baseline s-SIFT BoW method with χ^2 kernel attains a solid 90% accuracy on the microconstituent classification task, somewhat higher than the accuracy of $83 \pm 3\%$ recently reported for the same method applied to a much smaller (but more diverse) analogous 7-class microstructure classification task[1]. Switching to dense SIFT features slightly decreases the classification accuracy, especially for the cropped image set and with a smaller dictionary size.

VLAD-encoding improves the average SIFT-based image representation performance by up to an additional 6% for the microconstituent classification task, compared to the χ^2 BoW method. In the future, it may be interesting to explore competitive alternative normalization schemes for VLAD and Fisher encoding (i.e. the power normalization method[45], and applying the Hellinger kernel and PCA to individual SIFT features before the dictionary encoding step[46, 47, 48]).

The CNN-derived features we investigated consistently offer the best classification performance, though the marginal improvement of over VLAD-encoded SIFT features is roughly equal to the sample standard deviation. The marginal gain in classification performance results from moving from block4 features to higher-level CNN features is slight.

The performance differences between methods are much greater for the more difficult task of annealing condition classification. The variance in the performance estimate for any given method is also much higher for this task, primarily due to the very small dataset size. Again, the raw images yield classification accuracies that are substantially greater than the expected performance of a random classifier ($1/13 \simeq 7.7\%$). The SIFT-based BoW representations outperform the raw images by a much wider margin than for the microconstituent task. VLAD-encoding the SIFT features effectively doubles the accuracy compared to BoW encoding with the same dictionary size. It's clear from these results that the higher-level CNN features yield abstract microstructure representations that better capture to variations in annealing and imaging conditions, as further discussed in 3.2.

It is interesting to note that sparse SIFT representations yield classification accuracies as high or higher than dense SIFT representations. Speculatively, any potential benefit from a higher sampling density of local image features may be countered by the reduced rotation invariance of the fixed-orientation d-SIFT descriptors. For general microstructure characterization tasks, this rotation invariance is desirable, as any special orientation relationships (such as e.g. the preferred growth directions of Widmanstätten lath) are not necessarily related to the image reference frame.

Perhaps surprisingly, VGG_4 features seem to provide more discriminative representations for the spheroidite microstructures than the higher-level VGG_5 representations. Additionally, pooling multiscale VGG features seems plays no significant role on this task. One possible explanation is that the high-level convolution filters in the CNN are heavily optimized to perform an object recognition task, and provide discriminative abstract representations of objects in natural images–the fully-connected layers encode information about the global geometry of the objects detected in the image[40]. This can be mitigated by fine-tuning (re-training) the high-level CNN layers[49], or by employing additional feature pooling and encoding steps[39, 36, 50, 40], as we have attempted with VLAD encoding. An alternative possibility is that the present 3-class UHCS dataset is simply not a challenging classification task relative to current challenges in natural image recognition. Finally, variation in the labeling of the micrographs could also set an upper limit for classification accuracy on this dataset.

3.2 Data visualization

We used t-SNE to visualize the distributions of high-dimensional microstructure features obtained with each feature extraction method.[2] t-SNE is an unsupervised technique; the only input is the set of microstructure representations. Metadata such as primary microconstituent labels and processing metadata play no role on the structure of the resulting t-SNE maps. In Sections 3.2.1 and 3.2.2 we explore in more detail the image representation yielding the best classification results (VLAD-encoded VGG_5 features). Section 3.2.1 examines local variations in microstructure within the t-SNE representation, and Section 3.2.2 examines the relationship between microstructure features and the available processing metadata.

Bear in mind that t-SNE explicitly aims to reveal local structure within high-dimensional data, and that distances in the high-dimensional feature space can not always be exactly preserved in the low-dimensional t-SNE map[18]. As a result, large distances in the low-dimensional representation do not necessarily indicate large distances in the high-dimensional data, and high-distance 'seams' in the low-dimensional representation may exist where it is not possible to cleanly project the high-dimensional data. However, small distances in the t-SNE representation should correspond to a high degree of visual similarity.

[2]Because of space constraints, t-SNE maps and properties plots for many of the image representations investigated in this report were omitted in the manuscript. They are available in the supplementary materials and/or upon request.

3.2.1 t-SNE microstructure maps

In order to better understand how the microstructural representations are clustered and related in high dimensional space, we map them in two dimensions via the t-SNE visualization method. Figure 2 shows the t-SNE map for the entire UHCS dataset of 961 full-size images with markers color coded by primary microstructural constituent; the image representation is the VGG-block5 encoding, which was found to have the highest accuracy in the classification task. As shown in Figure 2, the spheroidite-related, pearlite-related, and network micrographs each form distinct, extended clusters, with subclusters denoting more closely related images, as discussed below. The martensitic images form two separate clusters that are closely related to the pearlite structures, as might be expected due to their similar aligned morphologies.The pearlite+widmanstatten images are more loosely clustered among the other pearlite-related images. A few outlier points in each category indicate micrographs that may challenge the image representation scheme, or may result from noise in the manual primary microconstituent process.

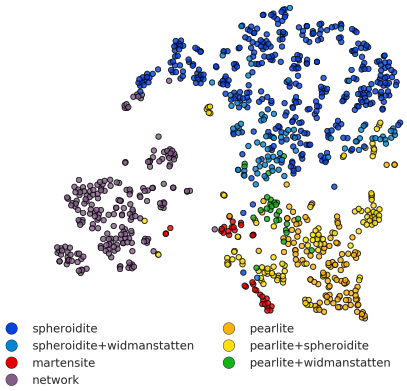


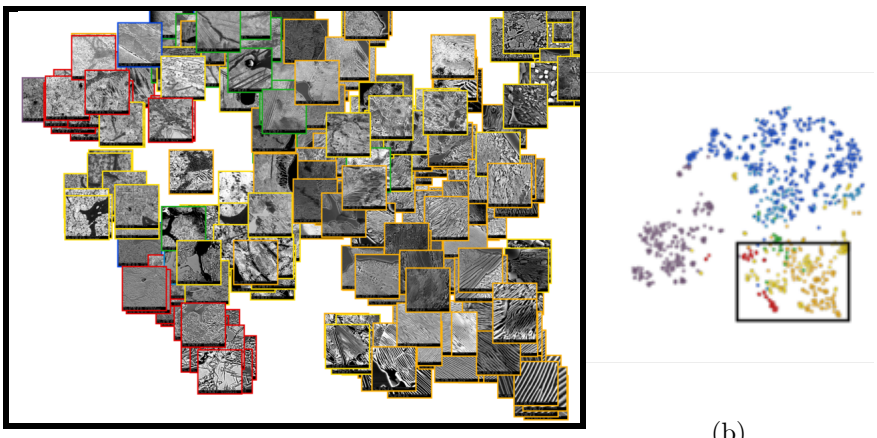
Figure 2: The VGG_5 $VLAD_{32}$ t-SNE map annotated with primary microconstituent.

Figures 3, 4, and 5 show detailed microstructure maps obtained by displaying each micrograph centered on its corresponding position in the VGG-block5 t-SNE map shown in Figure 2.[3] The black frames on the inset t-SNE scatter plots indicate which portion of the map is displayed. Colored

[3]The full microstructure map is much too large to display in format of the present paper, but is available in the supplemental materials, along with maps for other microstructure representations.

frames around each thumbnail image indicate the primary microconstituent label, following the same color map as used in Figure 2.

Figure 3a focuses on the visual appearance of the high-magnification pearlite micrographs, as indicated by in Figure 3b, as well as two apparent clusters of martensite structures at the left. These pearlite micrographs span multiple orientations, magnification, and lamellar spacing, increasing in complexity from the lower right corner of the map. The pearlite micrographs in the lower right corner of the map are high-magnification views of individual pearlite domains; traversing up to the top of the figure widens the field of view with more morphological variation, with clear trend in the lamellar orientation. The pearlite matrix often contains spheroidite in the micrographs in the top half of this figure.

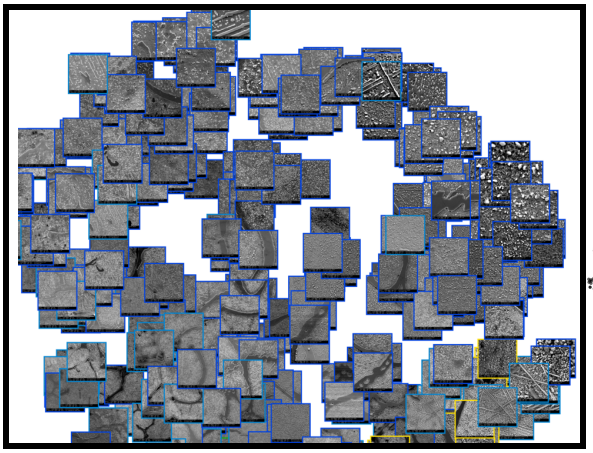


(a) (b)

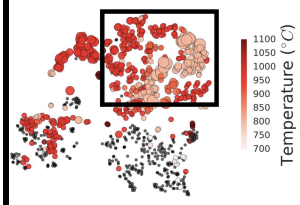
Figure 3: (a) VGG-block5 t-SNE microstructure map excerpt showing micrographs containing pearlite (b) in the lower-right portion of the VGG-block5 t-SNE map with markers coded by primary microconstituent (Figure 2).

Figure 4a shows most of the spheroidite micrographs in the block5 t-SNE map. These micrographs tend to cluster together with other micrographs from the same sample, i.e. the same processing conditions, as shown in Figure 4b. The magnification of these micrographs generally increases from the lower-left to the upper right quadrant of this map as well. Micrographs near at bottom of this map focus on the spheroidite-free denuded zones adjacent to the carbide network, while the higher-magnification micrographs at the top focus on the morphology and spatial distribution of individual spherodite particles. The microstructures in the upper left portion of this map are the

result of higher annealing temperatures compared to the micrographs in the lower right quadrant (see Figure 4b); this temperature gradient corresponds to a gradient in the spheroidite morphology across the map. One notable exception to this trend is the cluster of high-temperature microstructures in the lower-right corner of Figure 4a, which contain prominent Widmanstätten lath. While most of the material in this dataset was cooled by quenching, this cluster of micrographs come from material that was furnace-cooled.



(a)

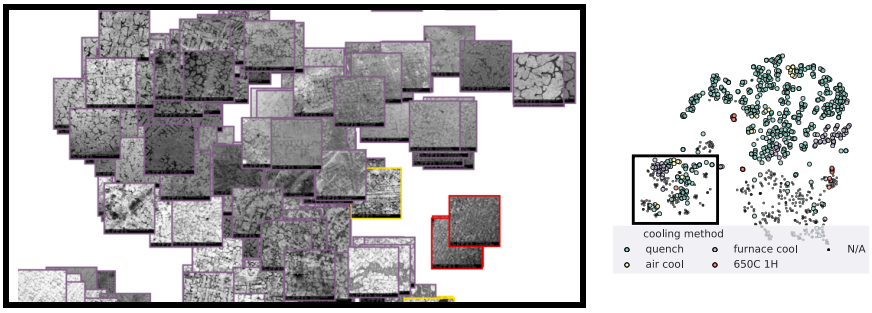


(b)

Figure 4: (a) VGG-block5 t-SNE microstructure map excerpt showing micrographs containing spheroidized cementite of various morphologies (b) in the left-most portion of the VGG-block5 t-SNE map, with marker colors indicating annealing temperature and relative marker sizes indicating annealing time (Figure 6e).

Figure 5a shows one the main cluster of proeutectoid cementite network microstructures. Many of these micrographs were subjected to similar processing conditions, with the bulk of them coming from samples annealed at $970°C$ for 90 minutes before being either air-cooled or water-quenched (see Figure 5b). The network structures formed under these common annealing conditions form two distinct groups in the VGG-block5 t-SNE map, in the upper left and central regions of Figure 5. The group at the upper left were furnace-cooled, while the central group were quenched. The quenched network micrographs clearly have lower contrast than the furnace-cooled network micrographs, and the cementite in the pearlitic matrix has a somewhat different morphology.

With its ability to sensibly arrange high-dimensional image representations



(a) (b)

Figure 5: (a) VGG-block5 t-SNE microstructure map excerpt showing micrographs containing high-level views of the proeutectoid cementite network (b) in the lower-right quadrant of the VGG-block5 t-SNE map with markers colored by cooling method (Figure 6d).

in a two-dimensional map, t-SNE is a valuable visualization method for microstructural image datasets, and for understanding the efficacy of high-dimensional microstructure representations. It enables quick visual scans to identify related images at large and small scales, captures systematic trends in microstructural morphology, and when coupled with processing metadata can graphically display the microstructure - processing link, as we discuss further in the next section.

3.2.2 Processing metadata

Though a regression model relating microstructural outcomes back to processing variables would be a more relevant model for a microstructure design task, the present dataset has an unbalanced distribution of processing parameters. However, examining these processing parameters by microstructural category still yields quantitative insight into the ability of the computer vision approach to infer processing - microstructure relationships. To this end, we map processing metadata onto the t-SNE map for image representations and explore the systematic trends between structure and processing.

Figure 6 illustrates the relationships between the available annealing schedule metadata and the resulting microstructure as shown by the VGG_5 t-SNE map from Figure 2. Figure 6a shows the annealing temperature in $^{\circ}C$; Figure 6b shows the annealing time in minutes; Figure 6c shows

the magnification (in microns per pixel) on a logarithmic scale; Figure 6d shows the cooling rate; and Figure 6e jointly illustrates the annealing time (proportional to marker size) and temperature (indicated by the color map). The small black markers indicate micrographs for which no processing metadata is currently available; these are mostly the high-magnification pearlite matrix micrographs, along with a subset of the network micrographs.

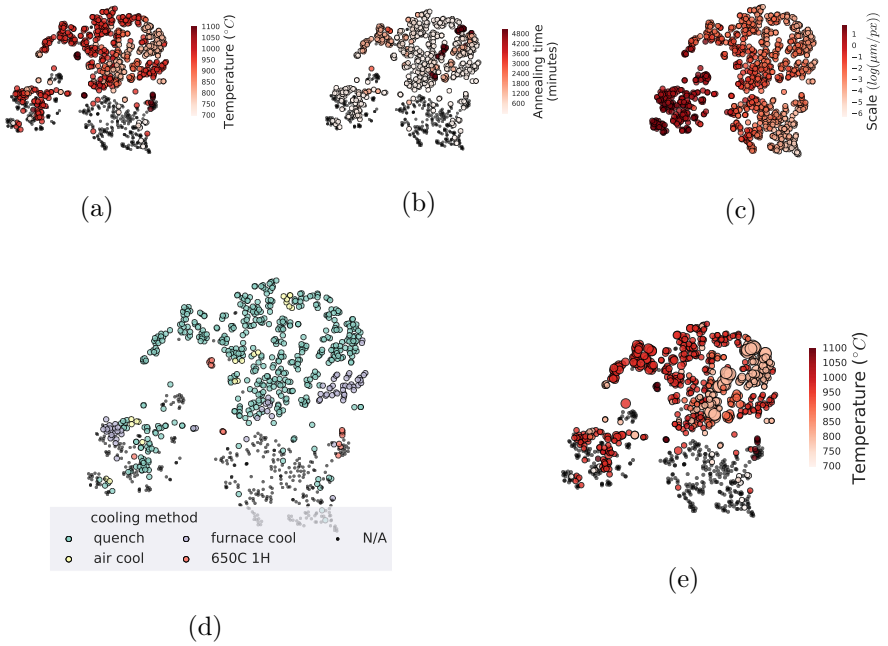


Figure 6: The VGG_5 $VLAD_{32}$ t-SNE map from Figure 2 annotated with processing metadata. Best viewed electronically. (a) Temperature (b) annealing time, (c) magnification in $\log_{10}(\mu m/px)$, (d) quench method, and (e) a bubble plot with colors indicating annealing temperature and marker size proportional to annealing time. For reference, the micrograph resolutions range from $0.002\mu m/px$ to $5.9\mu m/px$.

Apparent cluster structure in the VGG_5 t-SNE map clearly relates qualitatively to the annealing time and temperature data shown in Figure 6e. Most of the tightest local clusters in the t-SNE map consist of microstructures with the same or very similar annealing schedules. However, the magnification also plays a significant role. Consider the three small clusters of large, cream-colored points in the upper right quadrant of Figure 6e (low-temperature, long anneal micrographs of spheroidite) and their corresponding points in Figure

6c. The processing parameters and microstructures are similar between all three clusters, but the micrograph magnification increases by a factor of two moving from the upper right cluster to the middle cluster, and by yet another factor of two moving to the lower left cluster. The s-SIFT BoW and VLAD representation both also display this same effect; interestingly the VGG-pool5 representation seems to be somewhat more robust to changes in magnification, even though the scale-invariance of this method should be weaker and more implicit. Pooling the VGG_5 feature maps over multiple scales improves the situation (see the supplemental materials) by bringing the corresponding images to the upper two clusters closer together; the third cluster is still quite distinct, as those micrographs include substantially wider fields of view focusing on the carbide network and the morphology of the surrounding spheroidite. Thus, the question of how to incorporate the absolute physical scale of microstructure features into image representations adopted from the object and scene recognition communities must be addressed in order for these methods to help scientists and engineers develop quantitative processing–structure–properties mappings.

4 Conclusions

In this report, we establish a dataset for microstructure informatics that focuses on complex, hierarchical, and technologically-relevant microstructures. We evaluate applications of multiple image representation techniques from the field of computer vision in conjunction with both supervised and unsupervised microstructure informatics tasks. For this dataset, we show that appropriately pooled and encoded local features (SIFT) and domain-transferred deep convolutional neural network representations can provide classification accuracy better than 95%. We also discuss data visualization techniques (t-SNE) for exploratory analysis of microstructure and processing/properties metadata datasets. Explicit incorporation of the physical scale of microstructure features may be necessary for more quantitative microstructure science applications.

Acknowledgements

We gratefully acknowledge funding for this work through National Science Foundation grants DMR-1307138 and DMR-1501830, and through the John and Claire Bertucci Foundation. The UHCS micrographs were graciously provided by Matthew Hecht, Yoosuf Picard, and Bryan Webler (CMU). The

open source software projects VLFeat[51], Scikit-Learn[52], keras[53], and the reference implementation of t-SNE were essential to this work.

References

[1] Brian L DeCost and Elizabeth A Holm. A computer vision approach for automated analysis and classification of microstructural image data. *Computational Materials Science*, page in press, August 2015.

[2] Brian L. DeCost, Harshvardhan Jain, Anthony D. Rollett, and Elizabeth A. Holm. Computer vision and machine learning for autonomous characterization of am powder feedstocks. *JOM*, page Accepted for publication, March 2017.

[3] Aritra Chowdhury, Elizabeth Kautz, Bülent Yener, and Daniel Lewis. Image driven machine learning methods for microstructure recognition. *Computational Materials Science*, 123:176–187, 2016.

[4] Nicholas Lubbers, Turab Lookman, and Kipton Barros. Inferring low-dimensional microstructure representations using convolutional neural networks. *arXiv preprint arXiv:1611.02764*, 2016.

[5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[6] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[7] Tsung-Yu Lin and Subhransu Maji. Visualizing and understanding deep texture representations. *arXiv preprint arXiv:1511.05197*, 2015.

[8] Matthew D Hecht, Bryan A Webler, and Yoosuf N Picard. Digital image analysis to quantify carbide networks in ultrahigh carbon steels. *Materials Characterization*, 117:134–143, 2016.

[9] Matthew D Hecht, Yoosuf N Picard, and Bryan A Webler. Coarsening of inter and intragranular proeutectoid cementite in an initially pearlitic 2c-4cr ultrahigh carbon steel. *Metallurgical and Materials Transactions A*, Accepted for publication, 2017.

[10] Oleg D Sherby. Ultrahigh carbon steels, damascus steels and ancient blacksmiths. *ISIJ international*, 39(7):637–648, 1999.

[11] JM Hyzak and IM Bernstein. The role of microstructure on the strength and toughness of fully pearlitic steels. *Metallurgical Transactions A*, 7(8):1217–1224, 1976.

[12] J Pacyna and E Rożniata. Effect of annealing on structure and properties of ledeburitic cast steel. *Journal of Achievements in Materials and Manufacturing Engineering*, 24(1):84–90, 2007.

[13] Mingjia Wang, Songmei Mu, Feifei Sun, and Yan Wang. Influence of rare earth elements on microstructure and mechanical properties of cast high-speed steel rolls. *Journal of Rare Earths*, 25(4):490–494, 2007.

[14] KP Liu, XL Dun, JP Lai, and HS Liu. Effects of modification on microstructure and properties of ultrahigh carbon (1.9 wt.% c) steel. *Materials Science and Engineering: A*, 528(28):8263–8268, 2011.

[15] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2, 2004.

[16] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238, 2007.

[17] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[18] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[19] Jing Li and Nigel M Allinson. A comprehensive review of current local features for computer vision. *Neurocomputing*, 71(10):1771–1787, 2008.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[21] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 2015.

[22] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[23] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[24] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004.

[25] Stuart Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.

[26] Frederic Jurie and Bill Triggs. Creating efficient codebooks for visual recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 604–610. IEEE, 2005.

[27] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2559–2566. IEEE, 2010.

[28] Jan C Van Gemert, Cor J Veenman, Arnold WM Smeulders, and Jan-Mark Geusebroek. Visual word ambiguity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1271–1283, 2010.

[29] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[30] Piotr Koniusz, Fei Yan, and Krystian Mikolajczyk. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. *Computer vision and image understanding*, 117(5):479–492, 2013.

[31] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

[32] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.

[33] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013.

[34] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[35] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[36] Ali Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

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

[38] 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.

[39] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *Computer Vision–ECCV 2014*, pages 392–407. Springer, 2014.

[40] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3828–3836, 2015.

[41] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[42] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[43] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[44] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[45] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.

[46] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506. IEEE, 2004.

[47] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2911–2918. IEEE, 2012.

[48] Ken Chatfield, Victor S Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, volume 2, page 8, 2011.

[49] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(9):1904–1916, 2015.

[51] Andrea Vedaldi and Brian Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, pages 1469–1472. ACM, 2010.

[52] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[53] François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.