




# Describing Textures using Natural Language: Supplementary Material

Chenyun Wu , Mikayla Timm , and Subhransu Maji 

University of Massachusetts, Amherst  
{chenyun, mtimm, smaji}@cs.umass.edu

## 1 Dataset statistics and visualizations

We provide further details about the DTD<sup>2</sup>. Figure 1 shows the long-trail distribution of words and phrases in the dataset. Descriptions of categories such as “dots”, “lines”, “crystalline” and “checkers” are simpler than those of “spiralled”, “interlaced” and “crosshatched” as indicated by the number of unique words and phrases for each category. More examples of images and annotations from the dataset are shown in Figure 5, 6, and 7.

## 2 Comparison of ResNet101 features from different layers

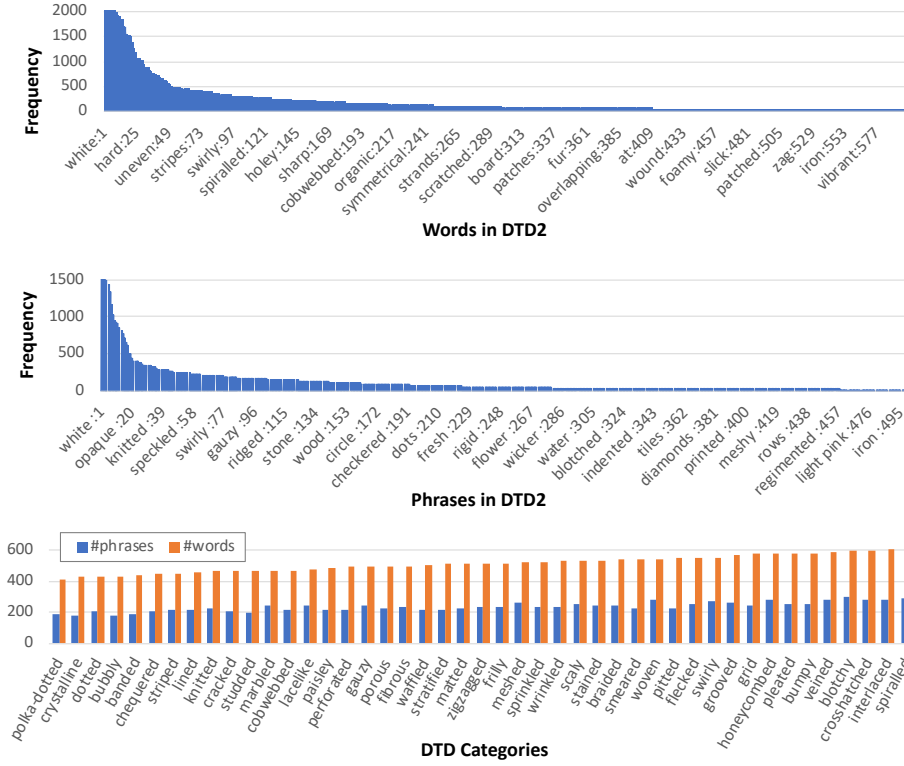
We compare features from layer-block 1 to 4 of ResNet101 in the classifier model, as shown in Table 1. Higher layer features perform better for phrase retrieval. For image retrieval, better performance is achieved with the combination of features from different layers. We select to use the features from layer 2 and 4 for all classifiers and metric learning models in all experiments with the classifier and the metric learning model.

## 3 Performance on different categories and phrases

Here we take a closer look at the performance of our models on different tasks on different phrases (Figure 2) and texture categories (Figure 3). The texture category is the category of the image in DTD. As seen in Figure 2-middle, the performance is correlated with the frequency of phrases, especially for image retrieval. As seen in Figure 3, we have better phrase retrieval performance on simpler categories that come with smaller vocabulary size.

## 4 Generated textures for analysis

Figure 4 shows the synthetically generated images and descriptions used in the experiments in Section 5.3. Each image is manually separated into two color regions which are systematically varied. In each row of the figure we systematically vary different attributes such as the foreground attribute, foreground

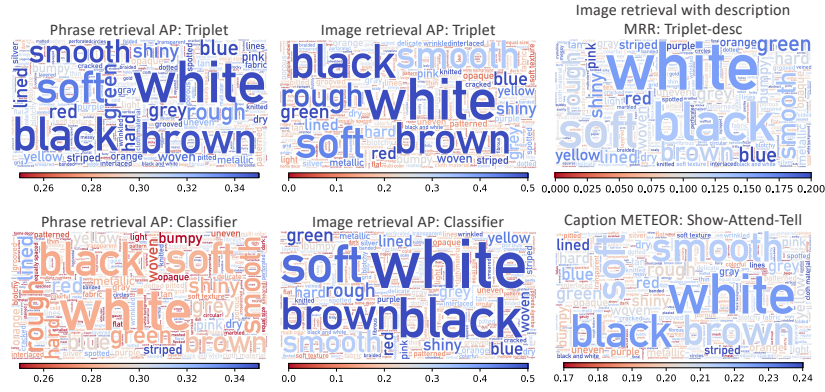


**Fig. 1. More Statics of DTD<sup>2</sup>.** On the top (middle) we show the histogram of word(phrase) frequency, where the x labels display some of the words(phrases) with their rankings in the dataset. On the bottom we show the number of unique words and phrases used to describe each DTD category, where we only count frequent words(phrases) that occur at least 5(10) times in training subset.

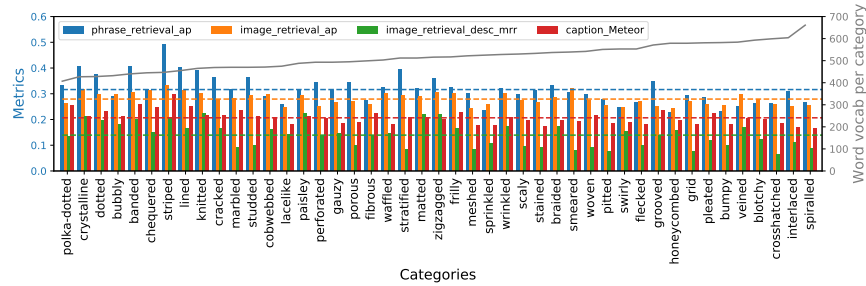
and background color, etc. This is a simple yet effective way to generate natural textures. Note how the texture properties are still maintained in each image, such as shading and winks of zebra stripes. This is much more challenging to do the same with natural images of scenes and objects.

**Table 1.** Performance on phrase retrieval and image retrieval on DTD<sup>2</sup>. “Classifier: Feat  $x$ ” stands for the classifier with image features from ResNet layer block  $x$  (or a concatenation of two layers.)

Data Split	Task:	Phrase Retrieval						Image Retrieval					
	Model	MAP	MRR	P@5	P@20	R@5	R@20	MAP	MRR	P@5	P@20	R@5	R@20
Validation	Classifier: Feat 1	13.10	37.20	16.05	10.68	4.94	13.04	10.64	25.06	11.57	9.37	6.13	17.78
	Classifier: Feat 2	17.65	44.91	22.41	14.59	6.85	17.60	13.00	29.24	14.60	11.08	8.52	22.54
	Classifier: Feat 3	26.43	<b>60.52</b>	32.47	20.71	9.93	25.00	15.62	31.79	17.28	13.34	9.42	28.52
	Classifier: Feat 4	26.51	59.24	<b>33.07</b>	20.84	<b>10.07</b>	25.16	15.85	<b>33.06</b>	17.83	13.02	9.94	27.28
	Classifier: Feat 1,4	25.78	58.28	31.58	20.31	9.55	24.44	15.85	32.35	<b>18.35</b>	13.51	10.24	28.03
	<b>Classifier: Feat 2,4</b>	26.57	59.19	32.65	21.11	9.99	25.50	<b>16.19</b>	32.53	17.47	<b>13.56</b>	<b>10.63</b>	<b>28.69</b>
	Classifier: Feat 3,4	<b>26.66</b>	60.38	32.20	<b>21.22</b>	9.81	<b>25.68</b>	16.04	31.18	17.59	13.50	10.33	28.32
Test	Classifier: Feat 2,4	27.12	61.28	33.50	21.71	16.07	41.48	<b>14.75</b>	<b>33.94</b>	<b>18.75</b>	<b>16.02</b>	<b>6.47</b>	<b>19.32</b>
	MetricLearning: BERT	<b>31.77</b>	<b>74.12</b>	<b>41.70</b>	<b>23.60</b>	<b>20.17</b>	<b>45.04</b>	13.50	31.12	16.52	14.57	5.24	17.32



**Fig. 2.** DTD<sup>2</sup> test set performance per phrase on each task with selected models. Color of phrases represents the metric performance: blue is better and red is worse, as indicated in the color bars. Font sizes are proportional to square root of phrase frequencies.

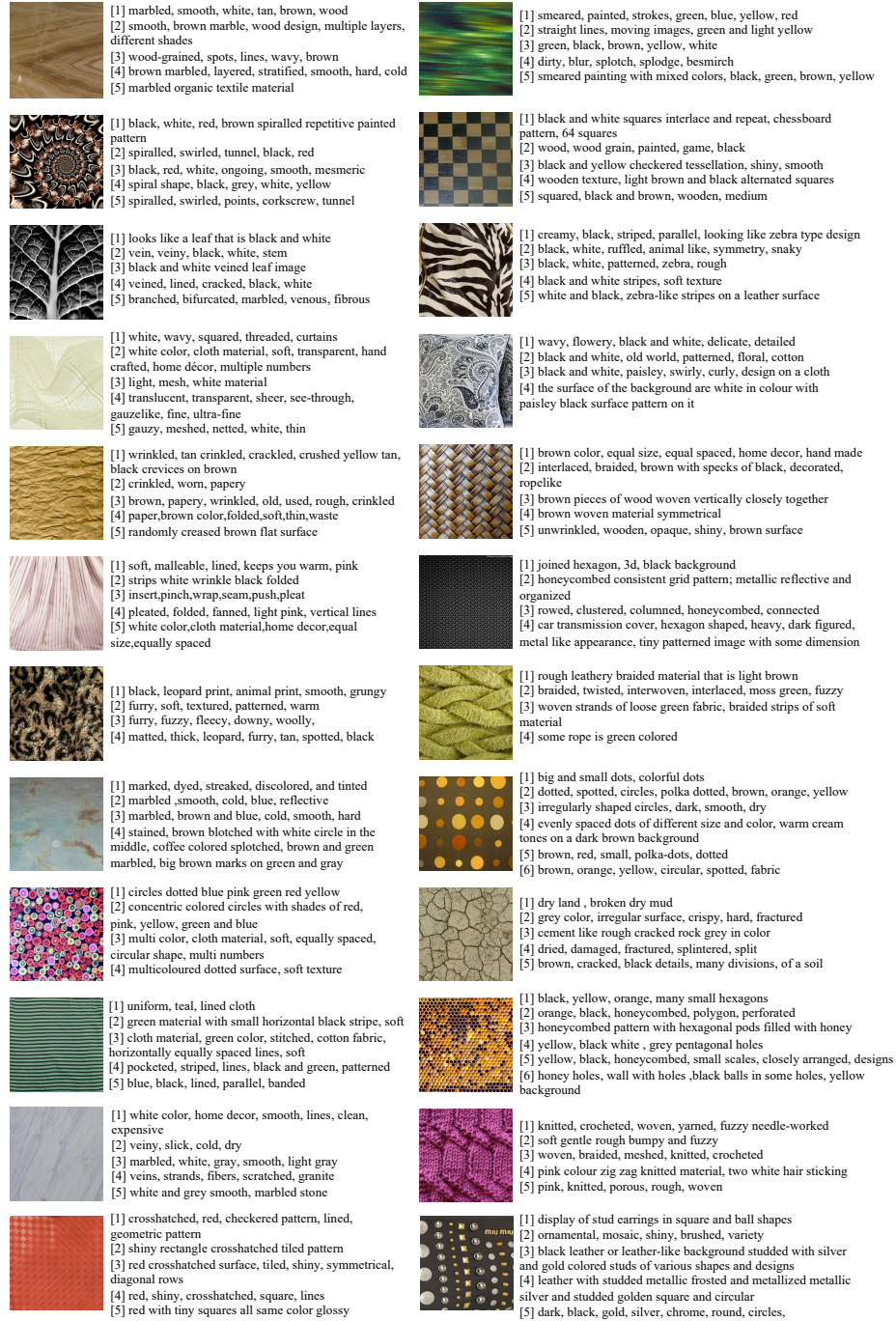


**Fig. 3.** DTD<sup>2</sup> test set performance per category on each task with our best model. Phrase retrieval: AP on our metric learning model; Image retrieval with phrase input: AP on our classifier; Image retrieval with description input: MRR on our metric learning model(description input); Captioning: METEOR on Show-Attend-Tell model.



**Fig. 4. Original and generated texture images.** The first row shows original images and their DTD categories. The rest rows are generated descriptions and images by sampling and modifying colors. The left 5 columns are images of Type A (foreground and background); the right 5 are Type B (no obvious foreground/background distinction).



Fig. 5. More examples from DTD<sup>2</sup>.

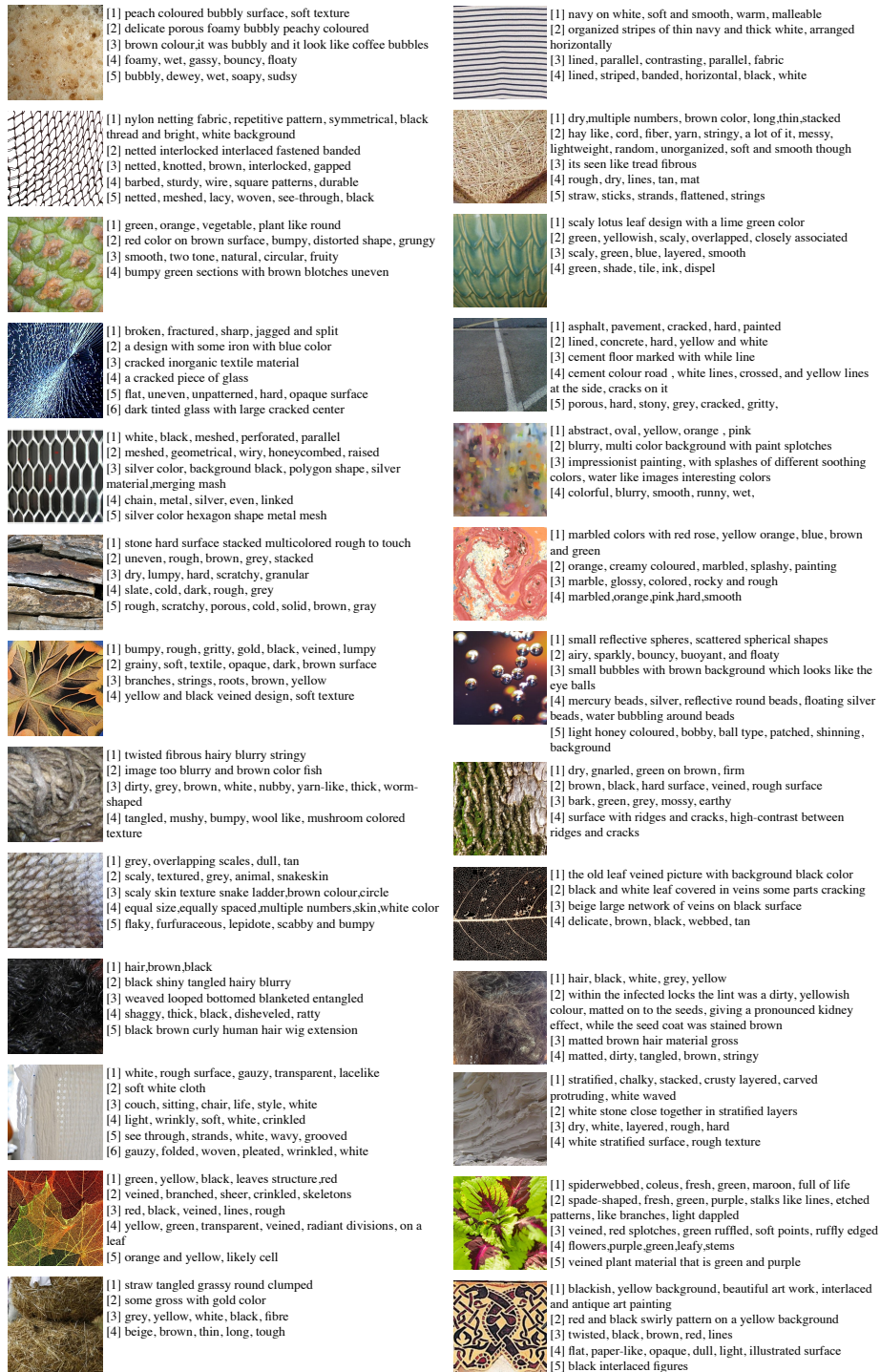
Fig. 6. More examples from DTD<sup>2</sup>.

Fig. 7. More examples from DTD<sup>2</sup>.