Enhancing Fine-Grained Classification for Low Resolution Images

Maneet Singh[1], Shruti Nagpal[1], Mayank Vatsa[2], and Richa Singh[2]

[1]IIIT-Delhi, India; [2]IIT Jodhpur, India

Email: {maneets, shrutin}@iiitd.ac.in, {mvatsa, richa}@iitj.ac.in

Abstract—Low resolution fine-grained classification has widespread applicability for applications where data is captured at a distance such as surveillance and mobile photography. While fine-grained classification with high resolution images has received significant attention, limited attention has been given to low resolution images. These images suffer from the inherent challenge of limited information content and the absence of fine details useful for sub-category classification. This results in low inter-class variations across samples of visually similar classes. In order to address these challenges, this research proposes a novel attribute-assisted loss, which utilizes ancillary information to learn discriminative features for classification. The proposed loss function enables a model to learn class-specific discriminative features, while incorporating attribute-level separability. Evaluation is performed on multiple datasets with different models, for four resolutions varying from 32×32 to 224×224. Different experiments demonstrate the efficacy of the proposed attribute-assisted loss for low resolution fine-grained classification.

*Index Terms*—Low Resolution Classification, Fine-grained, Attribute.

I. INTRODUCTION

Fine-grained classification focuses on distinguishing between objects belonging to distinct categories of the same parent class. Most of the existing research has focused on high resolution images [1]–[7], without much attention to low resolution data. Fine-grained classification of low resolution images is useful in surveillance scenarios, where low resolution data is captured from CCTV cameras. If a crime is captured on camera, clothes identification or vehicle make/model recognition is often important for obtaining supplementary information about suspects. The information can be circulated to create awareness or aid automated recognition. Other than surveillance, low resolution classification also finds applicability in data captured via smartphones containing objects of relatively smaller resolution. As observed in the literature, classification models trained on high resolution images often do not perform well on low resolution samples [8], thereby creating a need for algorithms designed specifically for low resolution fine-grained classification [9]–[11].

Fig. 1(a) presents sample high and low resolution images from the Animals with Attributes-2 [1] dataset. Low resolution samples are shown at 32×32 resolution, along with their corresponding bicubic interpolations. It is interesting to note that due to the similarity in structure and features, even at the high resolution of 128×128, it can be difficult to distinguish between the two images present in a row (such as Wolf and German Shepherd, or Wolf and Fox). In such cases,
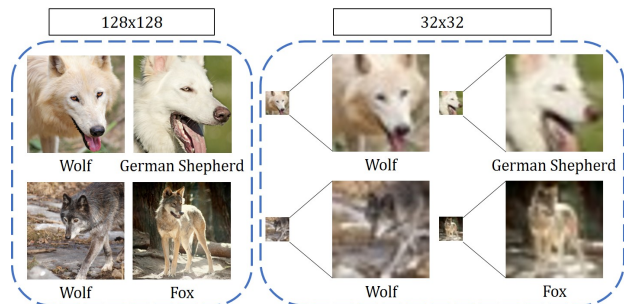
(a) High and low resolution samples from the AwA-2 dataset [1]
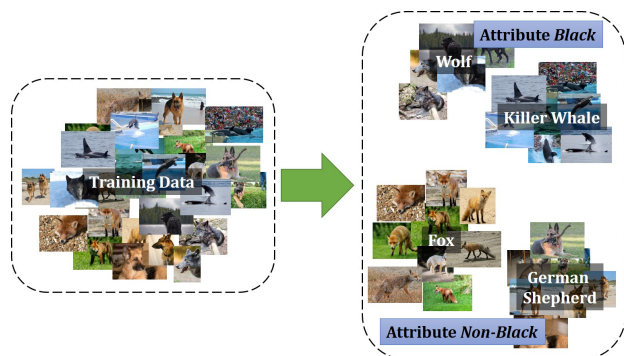
(b) Learned space via the proposed attribute-assisted loss

Fig. 1: (a) It is difficult to distinguish between data belonging to the same parent class, even at a high resolution; the problem is further aggravated in low resolution samples. (b) The proposed attribute-assisted loss introduces a higher level of separation based on the given attribute (*black*), followed by class-wise separability. Best viewed in color.

one generally utilizes fine details of localized regions such as the nose shape or eye color to supplement classification. However, at lower resolutions such as 32×32, these fine details are often lost, making the images almost indistinguishable. Therefore, in such scenarios, additional information which may or may not be present visually, can be utilized to aid recognition and enhance classification performance. Inspired by these observations, as shown in Fig. 1(b), this paper presents a novel *attribute-assisted loss* for low-resolution fine-grained classification. The proposed technique is built upon the philosophy of preventing attribute-level mis-classifications. For instance, if a low resolution image of a female subject is given as query, the output should be similar looking faces. In other words, a trained system should be able to differentiate
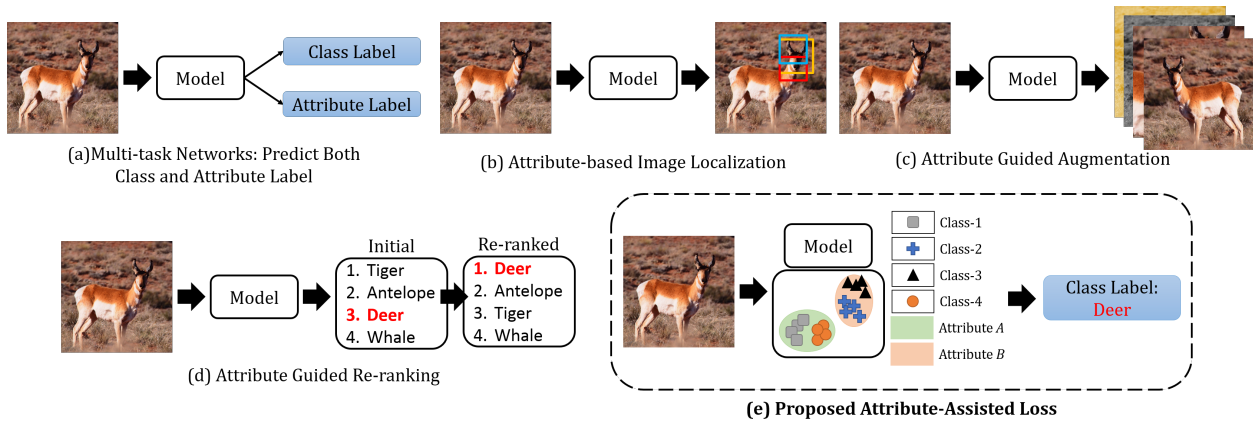
Fig. 2: (a)-(d): In the literature, researchers have utilized attribute information for fine-grained classification; (e): the proposed attribute-assisted approach. Different from existing techniques, the proposed loss implicitly utilizes the attributes to learn a more discriminative space, as opposed to predicting attributes or explicitly using them for ancillary tasks. Best viewed in color.

with respect to the inherent variations observed due to the gender attribute and return faces belonging to the 'female' attribute. Similarly, for the case of Wolf and German Shepherd, the model should clearly distinguish between *forest* and *non-forest* animals at a conceptual level[1]. Since attributes can provide a higher level separation between visually similar data, we propose incorporating them for enhancing fine-grained classification. Different from the existing techniques which explicitly utilize attributes for fine-grained classification, as shown in Fig. 2(a-d), the proposed attribute-assisted loss implicitly uses the attribute information for learning a better classification space. The contributions of this research are:

- We present a novel attribute-assisted loss for low resolution fine-grained classification. The proposed loss is model agnostic and utilizes attributes as ancillary information such that a network learns category-level discriminative features while incorporating attribute based similarity. To the best of our knowledge, this is the first research which utilizes attributes as ancillary information (which may or may not be visually present in the image) for aiding low resolution fine-grained classification.

- The effect of learning a network with both visual and non-visual (conceptual) attributes has also been studied in this research. Since the network learns an attribute-assisted discriminative feature space, improved performance is observed across both kinds of attributes.

- To the best of our knowledge, this is amongst the first few works which performs an in-depth evaluation for low resolution fine-grained classification, spanning across multiple resolutions of 32×32, 48×48, and 64×64, with three state-of-the-art deep learning models. Experiments on two publicly available datasets demonstrate the efficacy of the proposed loss.

II. RELATED WORK

Majority of the research in fine-grained classification has focused on high resolution images [12], with tremendous focus

[1]These attributes are provided in the Animals with Attributes-2 dataset.

on part detection, visual attention, and localization techniques [13]–[21]. Such methods isolate regions of the input image which contain distinctive characteristics useful for classification. For example, in case of bird classification, models are often trained to locate the wings, beak, or tail, in order to identify distinguishing features. Apart from attention or localization based networks, researchers have also proposed modifications to the Convolutional Neural Network architecture such as the Channel Max Pooling [22], Grassmann Pooling [23], Low-rank Bilinear Pooling [24], or Alpha Pooling [25], specifically for the task of fine-grained classification. Techniques which utilize additional information or encode features in a hierarchical manner have also shown to perform well [26], [27].

Low Resolution Fine-Grained Classification: The problem of low resolution fine-grained classification has received limited attention. In 2015, Chevalier *et al.* [11] studied the effect of decreasing the resolution of images for the task of fine-grained classification. They evaluated the performance of Fisher vectors and deep learning based representations for the given task. It was observed that the performance of fine-grained classification degrades rapidly below the resolution of 50×50. As demonstrated in Fig. 1(a), low resolution images often lack fine details, thereby rendering the task of part detection or localization challenging. In order to overcome this challenge, Cai *et al.* [10] proposed an end-to-end Convolutional Neural Network (CNN) based model for classification. The authors demonstrate improved performance in comparison with other techniques, by introducing an additional step of super resolution before classification. Often for smaller resolutions, the task of super-resolution is deemed challenging, since it requires synthesizing images of higher dimensions, with a large magnification factor, from limited information content.

Attributes for Fine-Grained Classification: As shown in Fig. 2, attribute information has widely been used for fine-grained classification for high resolution images. The existing techniques can broadly be categorized as follows: (a) multi-task networks [28], [29]; (b) algorithms which utilize attribute-
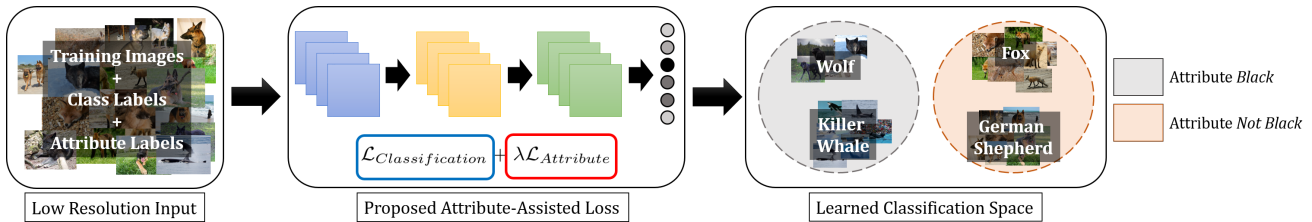
Fig. 3: Proposed attribute-assisted loss facilitates learning of a discriminative space for fine-grained classification. A higher level separability is enforced between classes having different attributes (eg. *black* and *not black*), while the trained model is able to distinguish between classes of the same attribute as well (eg. *Fox* and *German Shephard*). Images have been taken from the AwA-2 dataset [1].

based image localization, i.e. training networks *where* to look in an image [16], [27], [30]; (c) attribute-guided augmentation; or (d) attribute guided re-ranking of outputs [31], [32]. Most of these techniques either explicitly incorporate the visual attribute information in the classification network, or utilize an attribute prediction component. While this is feasible for high resolution images with high information content, we believe it is challenging for low resolution data. Therefore, in this work, we propose to utilize attributes as additional information, capable of assisting the network for discriminative features.

III. PROPOSED ATTRIBUTE-ASSISTED LOSS

Fig. 3 presents a broad overview of the proposed attribute-assisted loss, applicable to any *representation learning* algorithm. The proposed loss utilizes attribute information at the time of training in order to provide a higher-level of discrimination for low resolution fine-grained classification. Once the model is trained, it is used for classification on the test images, without requiring any attribute information. The proposed technique contains two components: (i) classification loss, and (ii) attribute loss, mathematically expressed as:

$$\mathcal{L}_{Proposed} = \mathcal{L}_{Classification} + \lambda \mathcal{L}_{Attribute} \qquad (1)$$

where, λ denotes the weight given to the attribute loss. The classification loss is used for learning the discriminative class-specific boundaries, while the attribute loss incorporates a higher level of separability between classes. The proposed attribute-assisted loss utilizes a triplet of training data: $\{x_i, y_i, a_i\}$, where x_i refers to the ith training image, y_i refers to its class label, and a_i refers to the attribute label of the ith sample. For n training samples, the proposed *attribute-assisted loss* is mathematically formulated as:

$$\mathcal{L}_{Proposed} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(p_{y_i}) \;+\; \lambda \frac{1}{2}\sum_{i=1}^{n} \|f(x_i) - r(a_i)\|_2^2 \qquad (2)$$

where, y_i is the ground-truth class label for the ith sample, and p_{y_i} refers to the prediction given by the model (a scalar value). $f(x_i)$ refers to the representation learned by the model for x_i (often of length 512 or 1024), and $r(a_i)$ refers to the representative feature for the scalar attribute a_i. In the above Equation, the first term corresponds to the cross-entropy loss used for classification, while the second term corresponds to the additional proposed attribute loss, which reduces the

intra-class variations with respect to the attribute information. While models trained with the cross-entropy loss only are useful for classification, however, in the scenario of low resolution fine-grained classification, they might not be able to learn meaningful representations due to the inherent lack of discriminative information content. It is our assertion that attribute information introduced via an additional attribute loss can help in learning better models, thus improving the classification performance.

For low resolution classification, where the model works with images having limited information, the additional attribute loss presents a two-fold purpose: (i) attribute labels are provided as additional information during training, thus increasing the available information for learning a discriminative model, (ii) the additional information facilitates learning of attribute encoded features such that the model does not confuse between samples being *conceptually* different; or in case of visual attributes, it forces the network to learn representations capable of discriminating on the given label. Considering the example of animal classification in low resolution images, the recognition framework should not confuse between animals found in *forests* with those found in *water*, and between animals *with* or *without patches*. The attribute loss facilitates feature learning such that the data corresponding to the same attribute label is brought closer, thus reducing the attribute-level mis-classifications. For an input x_i, this is achieved by minimizing the distance between the learned representation ($f(x_i)$) and the representative feature of its attribute, calculated as the mean feature of all samples with attribute a_i [33]. Mathematically, the attribute loss is formulated as:

$$\mathcal{L}_{Attribute} = \frac{1}{2}\sum_{i=1}^{n} \|f(x_i) - r(a_i)\|_2^2; \; r(a_i) = \mu(f(\mathbf{X}_{a_i})) \qquad (3)$$

where, $\mathbf{X}_{a_i}$ refers to all training samples having attribute a_i, and μ refers to the mean operator. Since the attribute loss is in conjunction with a classification loss, this enables the network to learn class-specific discriminative features, while incorporating the attribute-level separability as well. Attribute information is provided as input at the training time, and is thus not predicted from the network. This enables the model to learn attribute-assisted features useful for classification, while eliminating the need for attribute labels during testing.
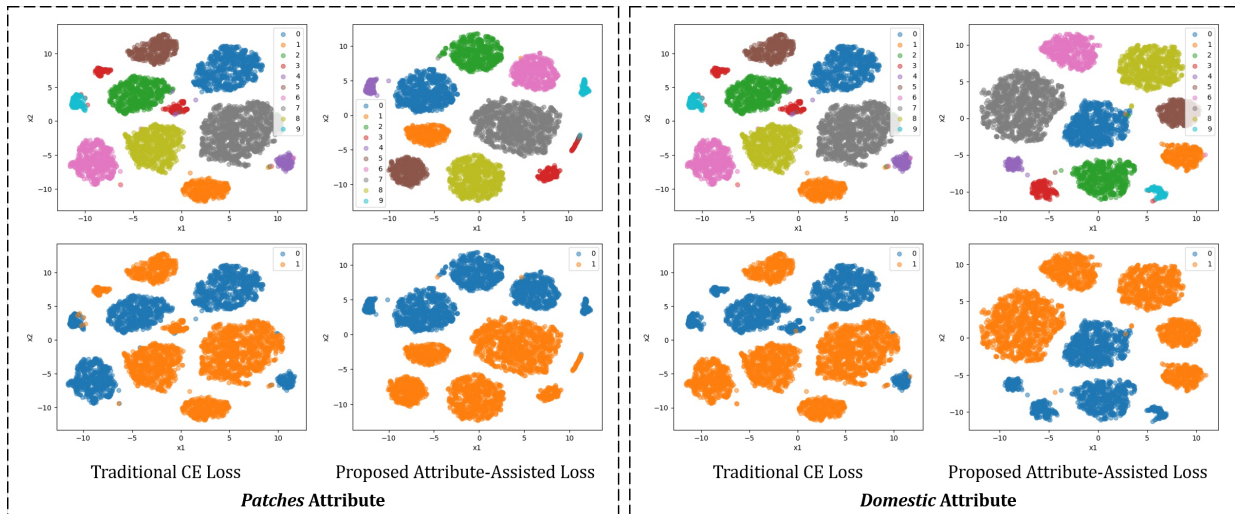
Fig. 4: t-SNE plot of learned features on a subset of the Animals with Attributes-2 dataset, containing 10 classes. The first row corresponds to the distribution of the learned representation with respect to the class labels, and the second row provides the attribute labels of the data points. It can be observed that attribute-assisted loss forces samples with the same attribute to be together, while the classification term enforces discrimination between the classes. Best viewed in color.

(a) Animals with Attributes-2 Dataset   (b) DeepFashion Dataset

Fig. 5: Sample images from the datasets used for experiments, for three different resolutions.

As a toy example, the benefit of learning a network via the attribute-assisted loss has visually been presented in Fig. 4. A subset of Animals with Attributes-2 (AwA-2) dataset [1] containing 10 classes is randomly chosen for this experiment, where 70% of each class is used for training, while the remaining samples form the test set. Pre-trained ResNet-18 [34] is trained using the traditional Cross Entropy (CE) loss and the proposed attribute-assisted loss for 32×32 resolution images. Visualizations are shown for two attributes: *patches* and *domestic*. The figure shows the final layer representations, projected onto a 2-D space using t-SNE [35]. It is interesting to observe that while both the models are able to discriminate well between the 10 classes, the models learned via the attribute-assisted loss are also able to differentiate the features based on the attribute. Classes belonging to a particular attribute are grouped together, while ensuring separability at the class-level. In this example, the attribute-assisted loss demonstrates an improvement of over 5% on the test set. This shows the benefits of the attribute-assisted loss for low resolution fine-grained classification.

As shown in Fig. 3, a model trained with the proposed loss utilizes the image samples, class labels, and the attribute labels while training. Model optimization is performed on Eq. 2 for

learning discriminative features, which encode the attribute information as well. During testing, the learned network utilizes only the input image, to predict the class label. Ground truth attribute label is not required for testing, nor does the model predict the attribute of the input sample.

IV. EXPERIMENTS AND IMPLEMENTATION DETAILS

The proposed approach is evaluated on two publicly available datasets: Animals with Attributes-2 [1] and DeepFashion [3] (Fig. 5). The details of these datasets are:

AwA-2 dataset [1] contains 37,322 images of animals belonging to 50 categories. Due to the lack of an existing protocol for fine-grained classification, we define the training and testing partitions. 70% data from each class is randomly chosen to form the training set, and the remaining samples form the test set. One sample of each class from the training set is randomly selected to create the validation set. The dataset also contains 85 binary attributes for each class, such as *white*, *domestic*, and *patches*. During training, data augmentation has been performed by random horizontal flipping.

DeepFashion dataset [3] contains over 8,00,000 images of clothes captured in different settings with varying pose, illumination, and resolution. A pre-defined protocol is provided

for clothing categorization, containing 289,219 images of 46 classes. The testing and validation set contain 40,000 images each, and the training set contains the remaining 209,219 images. Each image has been annotated with 1,000 attributes, along with a bounding box. Each class is also annotated with an additional attribute of *upper-body*, *lower-body*, or *full-body*. Since the proposed technique utilizes class-level attributes, the image-specific information has not been used in the experiments.

The efficacy of the proposed attribute-assisted loss has been demonstrated by performing experiments on three different models: ResNet-18, ResNet-50 [34], and DenseNet-121 [36]. All three models have shown to achieve superior performance on the ImageNet dataset [37], which contains over 3.2 million images of different objects. To demonstrate the effectiveness of the proposed loss, results have been computed by three methods: (i) extracting features from the pre-trained network and classifying using Euclidean distance, (ii) fine-tuning the pre-trained network using the standard Cross Entropy loss and performing classification, (iii) fine-tuning the pre-trained network using the proposed attribute-assisted loss and performing classification. The models have been evaluated across multiple resolutions, specifically, 32×32, 48×48, 64×64, and 224×224. The first two resolutions enable us to analyze the effect of the proposed model for low resolution classification, 64×64 enables us to understand its performance for moderate sized images, while 224×224 allows us to study the effect of the proposed loss for higher resolutions as well.

A. Implementation Details

The proposed attribute-assisted loss is implemented in PyTorch [38]. All pre-trained models have been taken from the Torchvision library and have been trained and tested using a Nvidia 1080Ti GPU. Each model is trained with the Stochastic Gradient Descent optimizer [39], with an initial learning rate of 0.01 and momentum of 0.9 for 100 epochs. The learning rate is reduced by a factor of 10 after every 20 iterations. A weight decay of 0.01 is also used as a regularizer during the training process. The weight of the attribute-assisted term (λ) is set to 0.01 after performing empirical evaluation. For smaller resolutions (32×32 and 48×48), a batch size of 400 samples is used for training, while for larger resolutions of 64×64 and 224×224, a batch size of 100 and 50 is used, respectively. For each experiment, all images have been synthetically down-sampled to the required resolution.

V. RESULTS AND ANALYSIS

The proposed attribute-assisted loss has been evaluated with three models and four resolutions. Results have been reported on the respective test set of the two datasets. Table I and Table II present the classification accuracies[2] obtained by the proposed attribute-assisted loss on the AwA-2 and DeepFashion datasets, respectively. For the proposed loss, *white* and *clothing type (upper, lower, or full body)* attributes have been

[2]Consistent with literature, rank-1 or top-1 accuracy has been reported, i.e. the percentage of samples correctly classified at the first rank only.

TABLE I: Rank-1 (top-1) classification accuracy obtained on the AwA-2 dataset for different resolutions and models. *Euc.* corresponds to the accuracy obtained upon matching the features obtained directly from the pre-trained model with Euclidean distance. *CE Only* refers to fine-tuning the pre-trained model with the traditional Cross Entropy loss, and *Proposed* refers to the accuracy obtained with the proposed attribute-assisted loss.

Resolution	Model	Euc.	CE Only	Proposed
32×32	ResNet-18	18.89	36.31	44.54
	ResNet-50	20.70	38.09	**53.24**
	DenseNet	18.98	41.36	50.56
48×48	ResNet-18	27.34	51.60	61.35
	ResNet-50	33.08	60.63	**67.10**
	DenseNet	28.09	61.59	64.29
64×64	ResNet-18	40.03	62.44	67.47
	ResNet-50	48.21	72.90	75.84
	DenseNet	44.55	75.94	**76.87**
224×224	ResNet-18	86.12	90.29	91.95
	ResNet-50	90.14	93.40	93.68
	DenseNet	85.51	93.57	**93.73**

TABLE II: Rank-1 (top-1) classification accuracy on the DeepFashion dataset for different resolutions and models. *Euc.* refers to the accuracy on matching the features obtained directly from the pre-trained model with Euclidean distance. *CE Only* refers to fine-tuning the pre-trained model with the traditional Cross Entropy loss, and *Proposed* refers to the accuracy obtained with the proposed attribute-assisted loss.

Resolution	Model	Euc.	CE Only	Proposed
32×32	ResNet-18	34.96	58.64	63.97
	ResNet-50	35.56	59.08	**64.02**
	DenseNet	36.36	60.98	62.49
48×48	ResNet-18	39.79	64.22	67.34
	ResNet-50	38.04	66.28	**67.96**
	DenseNet	40.02	64.69	67.41
64×64	Resnet-18	42.86	65.42	68.76
	ResNet-50	42.07	67.29	68.74
	DenseNet	44.02	67.90	**70.49**
224×224	Resnet-18	53.66	71.25	72.51
	ResNet-50	56.33	72.50	74.71
	DenseNet	56.62	72.46	**74.94**

used for the two datasets, respectively. The tables also present the accuracies obtained upon performing classification with the features extracted via the pre-trained model only with Euclidean distance (Euc.), and the accuracies obtained upon fine-tuning the pre-trained model with only the traditional Cross Entropy loss (CE Only). It can be observed that the proposed loss consistently improves the classification performance for the three models, across resolutions.

A. Effect of Resolution

Table I and Table II can be analyzed for resolution specific performance of the proposed attribute-assisted loss. On the AwA-2 dataset, a classification accuracy of 53.24%, 67.10%, 76.87%, and 93.73% is obtained for 32×32, 48×48, 64×64, and 224×224 resolutions, respectively. On the other hand, for the DeepFashion dataset, an accuracy of 64.02%, 67.96%, 70.49%, and 74.94% is obtained for the four resolutions,

respectively. For smaller resolutions of 32×32 and 48×48, the proposed attribute-assisted loss consistently performs best with the ResNet-50 model, while on higher resolutions, the DenseNet-121 model performs best.

The two tables can also be read from right to left in order to understand the *ablation study* on the proposed loss. The proposed technique utilizes both, the classification loss and the attribute loss for learning the network. On the other hand, 'CE only' uses only the classification loss. Therefore, the improvement observed across the two columns is attributed to the additional attribute assistance provided during learning. Fig. 6 presents the improvement in accuracy observed for the best performing models for both the datasets, across the resolutions. For example, for the AwA-2 dataset, an improvement of almost 15% is observed at a smaller resolution of 32×32 with the ResNet-50 model, whereas an improvement of less than 1% is observed at a higher resolution of 224×224. Similarly, for DeepFashion dataset, lower resolution classification benefits more from the proposed attribute-assisted loss as compared to higher resolution classification. We believe that at lower resolutions, where the information content is limited, additional information helps in learning better features, which in turn improves the classification performance. On the other hand, at higher resolutions, since the information content is as it is high, the effect of the attribute-assisted loss is less pronounced.

It is also interesting to observe that the improvement in accuracy is more prominent for the AwA-2 dataset, as compared to the DeepFashion dataset. It is important to note that the number of training samples for the AwA-2 dataset and DeepFashion differ almost by a ratio of 8 (AwA-2 has only around 26K training samples, while DeepFashion has over 200K samples). It is our assertion that for the smaller dataset of AwA-2, the proposed attribute-assisted loss also acts as a regularizer, thus preventing the model from overfitting on the training data. This also results in higher improvement at lower resolutions, where less information content coupled with relatively less data renders the problem further challenging.

B. Comparison with Existing Algorithms

Since low resolution fine-grained classification has garnered limited attention in the research community, this is the first work presenting the baselines for low resolution classification on the AwA-2 and DeepFashion datasets. Traditionally, these datasets have mostly been used for reporting results for zero-shot or few-shot protocols. In the literature, research has focused on high resolution fine-grained classification on the DeepFashion dataset as well. In this experiment, to compare with existing techniques, results are obtained for the proposed attribute-assisted loss using Densenet-121 on 224×224 resolution using the protocol given in Section IV. As per the protocol, Table IV presents the top-3 and top-5 accuracy of the proposed model along with other comparative techniques on DeepFashion database[3]. It can be observed that the proposed loss achieves the best performance by demonstrating an

---

[3]Consistent with literature, top-3 and top-5 accuracy has been reported, i.e. percentage of correct classifications within the top 3 or 5 ranks, respectively.
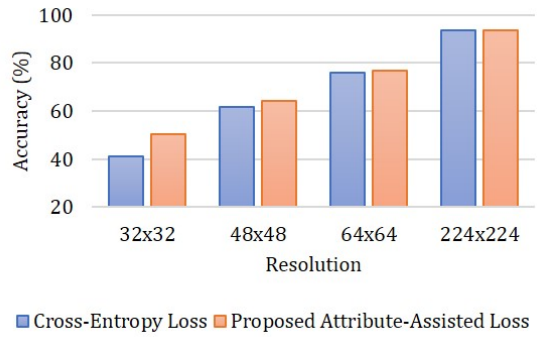
(a) Animals with Attributes 2
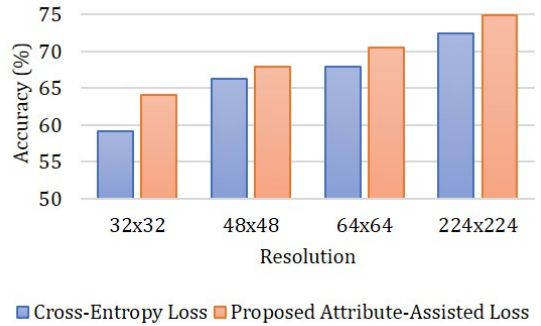
(b) Deep Fashion

Fig. 6: Comparison of classification accuracies of the best performing model for the two datasets. Accuracy obtained via the Cross Entropy loss is compared with the proposed attribute-assisted loss. It is interesting to note that the effect of the attribute information is more prominent in low resolution images, as compared to the higher ones.

improvement of over 8% and 3% for the top-3 and top-5 accuracies, respectively. One of the recent comparative techniques on this dataset, FashionNet, uses a CNN architecture to predict the attribute, landmark, and category of the input image. On the other hand, the proposed attribute-assisted loss pushes the model to learn features separable on a given attribute. The benefit of providing ancillary information, as opposed to performing classification on the attribute, can be observed in the improved performance of the proposed loss. Further, the pre-defined protocol on the AwA-2 dataset is for zero-shot learning only, and thus there exist limited comparative results for classification. In 2020, Xiao *et al.* [44] proposed a novel pooling block which obtains 86.79% classification performance on the AwA-2 dataset. In comparison, the proposed attribute-assisted loss demonstrates an improvement of around 7% by obtaining 93.73% (Table IV).

C. Effect of λ

As mentioned previously, the proposed attribute-assisted loss contains two main components: (i) the classification loss and (ii) the attribute loss. The classification loss corresponds to the traditional Cross Entropy loss used for classifying

TABLE III: Rank-1 (top-1) classification accuracy on the Animals with Attributes-2 dataset for 32×32 and 48×48 resolutions. Results are demonstrated with four different attributes - both visual and non-visual. Comparison with the traditional Cross Entropy loss (CE only) demonstrate the effectiveness of the proposed loss.

Resolution	Model	CE Only	Proposed Attribute-Assisted Loss			
			White	Patches	Domestic	Forest
32×32	ResNet-18	36.31	44.54	43.94	44.25	48.63
	ResNet-50	38.09	**53.24**	**52.60**	**51.84**	**52.80**
	DenseNet-121	41.36	50.56	49.16	46.36	49.11
48×48	ResNet-18	51.60	61.35	60.72	59.63	57.36
	ResNet-50	60.63	**67.10**	66.22	64.06	65.80
	DenseNet-121	61.59	64.29	**66.83**	**67.01**	**66.52**

TABLE IV: Comparison of the proposed attribute-assisted loss on DenseNet-121 with state-of-the-art techniques on the DeepFashion dataset and the AwA-2 dataset for 224×224 resolution. Due to the same protocol, comparison has been performed with the ranks reported in the literature.

DeepFashion Dataset		
Algorithm	Top-3	Top-5
WTBI (2012) [40]	44.73	66.26
DARN (2015) [28]	59.48	79.58
FashionNet+Joints (2016) [3]	72.30	81.52
FashionNet+Poselets (2016) [3]	75.34	85.87
FashionNet (2016) [3]	82.58	90.17
Lu *et al.* (VGG-16) (2017) [41]	86.72	92.51
Corbiere *et al.* (2017) [42]	86.30	92.80
SEResNeXt50+OC (2019) [43]	88.42	93.93
Proposed Attribute-Assisted Loss	**91.05**	**95.35**
AwA-2 Dataset		
Algorithm	Top-1	
Xiao *et al.* (2020) [44]	86.79	
Proposed Attribute-Assisted Loss	**93.73**	

Fig. 7: Effect of λ, weight of the attribute-loss. Analysis is performed on 32×32 resolution, with ResNet-50 and DenseNet-121, for the AwA-2 dataset.

the input data into one of the classes. The attribute loss works with the ancillary information and attempts to minimize mis-classifications between classes of different attributes. As shown in Eq. 2, λ controls the contribution of the attribute-assisted loss in the final loss function. In this subsection, the contribution of the attribute-assisted loss is analyzed based on the weight parameter. Experiments have been performed on the AwA-2 dataset, with low resolution data (32×32). Fig. 7 demonstrates the effect of varying the weight parameter (λ) on the classification accuracy. Experiments have been performed on the top two models: Resnet-50 and Densenet-121. A similar trend is observed for both the models, where increasing the weight value results in an increase of the classification performance, till a particular value. At a weight of 0.01, both the models reach the peak performance, beyond which the performance of the model declines. These trends are intuitive in nature, for instance, at a smaller weight value, the contribution of the attribute-assisted term is less, whereas on larger weight values, the contribution of the attribute loss offsets the classification loss, thereby resulting in a drop in the performance.

D. Evaluation with Different Attributes

The proposed attribute-assisted loss has also been evaluated with different attributes. Since the DeepFashion dataset contains only a single set of cl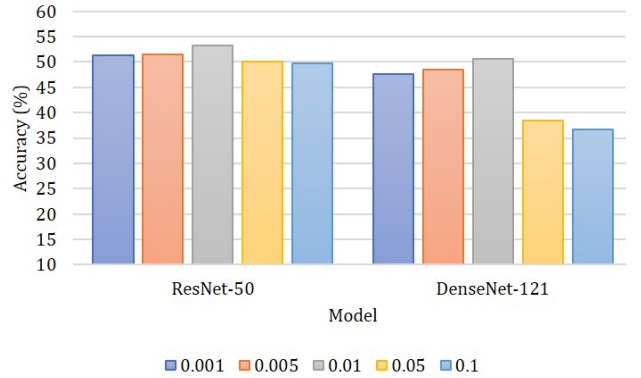ass-wise attributes, experiments are performed on the AwA-2 dataset only. Four different attributes of *white*, *patches*, *domestic*, and *forest* have been considered for evaluation, with lower resolutions of 32×32 and 48×48 resolutions. Table III presents the performance of the proposed attribute-assisted loss with different attributes, along with the performance of the traditional Cross Entropy loss (CE only). Table III shows that the proposed attribute-assisted loss is able to learn improved classifiers with different types of attributes. For example, while *white* and *patches* provide explicit visual information about the animal, *domestic* and *forest* can be viewed as non-visual cues. The proposed attribute-assisted loss is able to utilize both these categories of information for enhancing the performance.

VI. CONCLUSION

Fine-grained classification has garnered substantial attention over the past few years, with majority of the algorithms focusing on high resolution images only. This research proposes a novel attribute-assisted loss for the challenging problem of low resolution fine-grained classification. The proposed loss utilizes attributes as additional information, thereby facilitating a given model to learn attribute-level discriminative features while performing classification. One of the key highlights of the proposed loss is its architecture agnostic behavior, and the ability to incorporate non-visual attributes as well. The efficacy of the proposed attribute-assisted loss is demonstrated

with three networks (ResNet-18, ResNet-50, and DenseNet-121) on two publicly available datasets (DeepFashion and AwA-2 datasets), for four varying resolutions from 32×32 to 224×224. The improved performance obtained across different resolutions strengthens its usage for low resolution fine-grained classification. As future work, the proposed technique can be extended to handle multiple attributes during learning.

VII. Acknowledgement

References

[1] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learninga comprehensive evaluation of the good, the bad and the ugly," *IEEE T-PAMI*, vol. 41, no. 9, pp. 2251–2265, 2019.

[2] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *IEEE Workshop on 3dRR*, 2013, pp. 554–561.

[3] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *IEEE CVPR*, 2016, pp. 1096–1104.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[5] A. Pal, G. Krishnan, M. R. Moorthy, N. Yadav, A. R. Ganesh, and T. S. Sharmila, "DICENet: Fine-Grained Recognition via Dilated Iterative Contextual Encoding," in *IJCNN*, 2019.

[6] J. Jiang and H. Wang, "Multi-grained selection and fusion for fine-grained image representation," in *IJCNN*, 2020.

[7] A. Majumdar, R. Singh, and M. Vatsa, "Face verification via class sparsity based supervised encoding," *IEEE T-PAMI*, vol. 39, no. 6, pp. 1273–1280, 2017.

[8] M. Singh, S. Nagpal, M. Vatsa, R. Singh, and A. Majumdar, "Identity aware synthesis for cross resolution face recognition," in *IEEE CVPR Workshops*, 2018, pp. 479–488.

[9] H. S. Bhatt, R. Singh, M. Vatsa, and N. K. Ratha, "Improving cross-resolution face matching using ensemble-based co-transfer learning," *IEEE T-IP*, vol. 23, no. 12, pp. 5654–5669, 2014.

[10] D. Cai, K. Chen, Y. Qian, and J.-K. Kmrinen, "Convolutional low-resolution fine-grained classification," *PRL*, vol. 119, pp. 166 – 171, 2017.

[11] M. Chevalier, N. Thome, M. Cord, J. Fournier, G. Henaff, and E. Dusch, "LR-CNN for fine-grained classification with varying resolution," in *IEEE ICIP*, 2015, pp. 3101–3105.

[12] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *IJAC*, vol. 14, no. 2, pp. 119–135, 2017.

[13] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *IEEE CVPR*, 2017, pp. 4476–4484.

[14] X. He and Y. Peng, "Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification," in *AAAI*, 2017.

[15] K. Oh, S. Kim, and I.-S. Oh, "Salient explanation for fine-grained classification," *IEEE Access*, vol. 8, pp. 61 433–61 441, 2020.

[16] X. Liu, J. Wang, S. Wen, E. Ding, and Y. Lin, "Localizing by describing: Attribute-guided attention localization for fine-grained recognition," in *AAAI*, 2017, pp. 4190–4196.

[17] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE T-IP*, vol. 27, no. 3, pp. 1487–1500, 2018.

[18] H. Hanselmann and H. Ney, "ELoPE: Fine-grained visual classification with efficient localization, pooling and embedding," in *WACV*, 2020, pp. 1247–1256.

[19] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *ECCV*, 2018, pp. 420–435.

[20] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE T-MM*, vol. 19, no. 6, pp. 1245–1256, 2017.

[21] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *IEEE ICCV*, 2017, pp. 5219–5227.

[22] Z. Ma, D. Chang, J. Xie, Y. Ding, S. Wen, X. Li, Z. Si, and J. Guo, "Fine-grained vehicle classification with channel max pooling modified cnns," *IEEE T-VT*, vol. 68, no. 4, pp. 3224–3233, 2019.

[23] X. Wei, Y. Zhang, Y. Gong, J. Zhang, and N. Zheng, "Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification," in *ECCV*, 2018, pp. 355–370.

[24] S. Kong and C. C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *CVPR*, 2017, pp. 7025–7034.

[25] M. Simon, Y. Gao, T. Darrell, J. Denzler, and E. Rodner, "Generalized orderless pooling performs implicit salient matching," in *IEEE ICCV*, 2017, pp. 4960–4969.

[26] S. Cai, W. Zuo, and L. Zhang, "Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization," in *IEEE ICCV*, 2017, pp. 511–520.

[27] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," in *IEEE ICCV*, 2017, pp. 1349–1358.

[28] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *IEEE ICCV*, 2015, pp. 1062–1070.

[29] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *ICCV*, 2013, pp. 729–736.

[30] Y. Yan, B. Ni, and X. Yang, "Fine-grained recognition via attribute-guided attentive feature aggregation," in *ACM MM*, 2017, pp. 1032–1040.

[31] M. Dixit, R. Kwitt, M. Niethammer, and N. Vasconcelos, "Aga: Attribute-guided augmentation," in *IEEE CVPR*, 2017, pp. 7455–7463.

[32] S. Xie, T. Yang, X. Wang, and Y. Lin, "Hyper-class augmented and regularized deep learning for fine-grained image classification," in *IEEE CVPR*, 2015, pp. 2645–2654.

[33] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[35] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, pp. 2579–2605, 2008.

[36] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, 2017, pp. 2261–2269.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.

[38] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[39] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *COMPSTAT*, 2010, pp. 177–186.

[40] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *ECCV*, 2012, pp. 609–623.

[41] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," in *IEEE CVPR*, 2017, pp. 5334–5343.

[42] C. Corbiere, H. Ben-Younes, A. Ramé, and C. Ollion, "Leveraging weakly annotated data for fashion image retrieval and label prediction," in *IEEE ICCV*, 2017, pp. 2268–2274.

[43] S. Park, M. Shin, S. Ham, S. Choe, and Y. Kang, "Study on fashion image retrieval methods for efficient fashion visual search," in *IEEE CVPR Workshops*, 2019.

[44] B. Xiao, X.-y. Li, C.-g. Li, and Q.-f. Xu, "A novel pooling block for improving lightweight deep neural networks," *PRL*, vol. 135, pp. 307–312, 2020.