Authomatic dataset generation procedure

In this document we describe how we built the large-scale datasets of Double JPEG (DJPEG) pristine and synthetic tampered images used for the experiments of our paper 'Image Splicing Detection, Localization and Attribution via JPEG Primary Quantization Matrix Estimation and Clustering'. The reader may refer to the paper (available at BTcomm:....??? insert the link to the paper) for the notation.

We considered two types of DJPEG pristine and tampered images, named Type I and Type II, respectively for the case of Aligned DJPEG (A-DJPEG) and the case of Non-Aligned DJPEG (NA-DJPEG).

- For **Type I** images: the pristine images are A-DJPEG images. This means that the first 8×8 DCT compression grid is aligned with the grid of the second compression. With regard to the tampered images, the first JPEG compression grid of the background (or, equivalently, of the source image) is aligned with the grid of the second compression, while the grid for the first compression of the foreground is misaligned with that of the second compression. Specifically, a random misalignment (r, c), $0 \leq r, c \leq 7, (r, c) \neq (0, 0)$ is considered for the grid of the former compression of the foreground.
- For **Type II** images: we assume that the images are first JPEG compressed using a DCT grid shifted by a quantity (r, c), randomly chosen in $0 \leq r, c \leq 7, (r, c) \neq (0, 0)$, with respect to the upper left corner, while for the second compression no grid misalignment is considered. Then, the pristine images are NA-DJPEG. For the tampered images, the JPEG grid of the background is non-aligned with the grid of the second compression. The same holds for the foreground regions. Note that the misalignments of foreground and background are generally different.

To built the datasets for our experiments, we started from the 8156 camera-native uncompressed large size images in the RAISE8K dataset [1]. We divided these images in two sets: 7000 images to be used for training (and validation) and 1156 images for the tests. On the average, about 5 non-overlapping patches are extracted from each RAISE image, for a total number of 41000 patches, 35000 of which (coming from the set of 7000 original images), denoted by $\mathcal{S}_{tr}$, were used to produce the pristine and tampered images for training the models and the remaining 5780, denoted by $\mathcal{S}_{ts}$, to produce the pristine and tampered images for the tests.

We considered images of size 512×512[1]. These images are obtained from larger size uncompressed images by dividing them in non-overlapping patches. These patches are then processed to get the Type I and Type II images (the patches have size 528×528, so that a 512×512 crop can always be obtained

[1]This choice was made to speed up the computation of the $\widehat{Q}_1$ tensor, involving window-based estimation via a CNN (see the paper).

after the shift). To avoid picking up patches with a uniform content (e.g., sky patches), we put a control on the variance so that patches with a very small variance are discarded. Specifically, we set a threshold equal to 10 on the standard deviation. In the sequel, we will refer to this set of blocks of size 528×528 as $\mathcal{D}_p$.

Several QF_1 are considered for the first JPEG compression. The second JPEG quality factor QF_2 is fixed to 90 in our experiments. Such value corresponds to a common yet not too high quality, so that both the scenarios $QF_1 < QF_2$ and $QF_1 > QF_2$ make sense. For the tampered images, we set $k = 2, 3$ and 4; we denote with QF_1 the first quality factor for the background, while $\{QF_{1,i}\}_{i=2}^{k}$ denotes the first quality factors for the tampered regions (in the most general case of 3 different donors). Finally, we refer to the size $h \times w$ of the bounding-box of the spliced region as the tampering size.

Below, we describe how we built the pristine and tampered DJPEG images used in our experiments. We denote the images before tampering as *source* images.

Pristine images. For a given first quality factor QF_1, the pristine DJPEG compressed images are obtained as follows. The starting images are (randomly) chosen in the set of patches $\mathcal{D}_p$ and JPEG compressed with QF_1. For Type II images, a random (r, c) shift ($(r, c) \neq (0, 0)$) is applied to the $8\times$ 8 grid to get the misalignment (as detailed before). The pristine images are then obtained by cropping them to 512 $\times$ 512 (aligned crop with the original grid) and then compressing them with QF_2. This same procedure is used to generate the source JPEG images used for the tampering, with the exception of the last compression step (source images are single JPEG images with QF_1).

Tampered images. To generate the tampered images for a given k, $\{QF_{1,i}\}_{i=2}^{k}$, and tampering size, we proceed as follows. We generate the source JPEG compressed images with QF_1 as detailed above. The tampered region is obtained from another image, JPEG compressed with quality factor $QF_{1,2}$, randomly chosen from the same set of images (different from the one used for the source image), called *donor* image. The shape of the tampered region is a convex polygon randomly generated (from a subset of 20 random vertices, selected in such a way that they form a convex hull) within a bounding box of sizes $h \times w$, randomly located in the image. The pixels inside the convex polygon form the region that is going to be spliced. The random crop selected from the donor image is copy-pasted into the same position of the source image.

If $k = 2$, the tampered image obtained in this way is JPEG-compressed again with QF_2, and the procedure ends. If $k > 2$, another (different) donor image is chosen, JPEG compressed with $QF_{1,3}$, and the randomly selected polygon is copy-pasted in the target image. If the two tampered regions overlap, another random position for the bounding-box is selected (until no overlap occurs). Then, if $k = 3$, the tampering process ends and the image is JPEG compressed with QF_2, otherwise, the procedure is repeated

to perform the additional tampering with $QF_{1,4}$, followed by a final JPEG compression with QF_2. Fig. 1 illustrates the forgery creation process for a given $k = 3$.

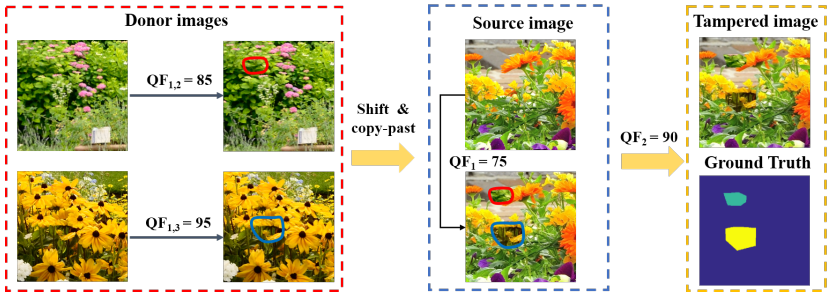

Fig. 1: Illustration of the tampering generation process with $k = 3$. Two donor images are compressed with $QF_{1,2}$ and $QF_{1,3}$, respectively, and then a random grid shift is applied to them. Two tampered regions are selected within two bounding boxes, of sizes 96×128 and 156×156 in the figure. The source image is compressed with $QF_1 = 75$ and then cropped to 512×512 (aligned or misaligned). The two tampered regions are copy-pasted in the source image compressed with QF_1 (aligned or not). Then, the tampered image is finally compressed with QF_2.

Below, we provide a description of the datasets considered for the various tests in our experiments. The pristine and tampered images are built following the above detailed procedure.

A. Dataset for Q_1 matrix estimation (training and testing)

To build the datasets for training and testing the CNN for the estimation of Q_1, we followed exactly [2]. Training and testing was carried out on 64×64 patches, obtained from the set of 7000 and 1156 images of RAISE. For DJPEG, a random grid shift (r, c) is applied between the two compressions, $0 \leq r, c \leq 7$, then, as in [2], the A-DJPEG case occurs with probability 1/64.

B. Dataset for k estimation (training and testing)

The dataset used to train and test the CNN for the estimation of the number of clusters consists of a total of 80000 images, 20000 for each k, that is, 20000 pristine images ($k = 1$) and 60000 tampered images ($k = 2, 3, 4$). The dataset consists of the following subsets:

- a set $\mathcal{D}_{tr}$ of 18000 images for each k (for a total of 72000 images) used for training. The set is obtained from 18000 (randomly chosen) images in $\mathcal{S}_{tr}$;
- a set $\mathcal{D}_{ts}$ consisting of 2000 images for each k (for a total of 8000 images) used for testing. The set is obtained from 2000 images in $\mathcal{S}_{ts}$.

Let $V = \{60, 65, 70, 75, 80, 85, 95, 98\}$. To get the pristine images, QF_1 is randomly chosen in V and $QF_2 = 90$. For the tampered images, when $k = 2$, QF_1 is randomly chosen in $\{75, 85, 95, 98\}$, and

$QF_{1,2} \in V$, $QF_{1,2} \neq QF_1$. When $k = 3$, $QF_1 \in \{75, 85, 95, 98\}$, and $QF_{1,2}, QF_{1,3} \in V$, $QF_{1,2} \neq QF_{1,3} \neq QF_1$. Finally, for $k = 4$, $QF_1 \in \{75, 85, 95, 98\}$, and $QF_{1,2} \neq QF_{1,3} \neq QF_{1,4} \neq QF_1 \in V$. The height h and width w of the bounding-box of the tampered regions are randomly selected in {64, 96, 128, 156} (see the Appendix for the details of the tampered image generation). Misalignment is applied to the background with 0.5 probability, then the dataset consists of both Type I and Type II images in similar proportions.

C. Dataset for detection, localization and attribution tests

Detection performance are measured over the same dataset $\mathcal{D}_{ts}$ considered to test the CNN for k estimation, where we have 2000 images representative of the negative class (pristine), and 6000 for the positive class (tampered). The threshold T necessary to achieve the desired FPR is set on these 2000 pristine images.

To better assess the localization performance, and to ease the comparison with state-of-the-art methods (see the next section), we additionally built two separate Type I and Type II datasets, named $\mathcal{D}_I$ and $\mathcal{D}_{II}$, whose images are generated from $\mathcal{S}_{ts}$ under specific setting. Specifically, in both $\mathcal{D}_I$ and $\mathcal{D}_{II}$, we considered 100 images for every combination of k and $\{QF_{1,i}\}_{i=2}^{k}$, for the tampering sizes $h \times w = 96 \times 96$ and 128×128.

A summary of the datasets used in our experiments is reported in Table I.

TABLE I: Datasets of images considered in our experiments.

Name	$\mathcal{D}_{\mathbf{tr}}$	$\mathcal{D}_{\mathbf{ts}}$	$\mathcal{D}_{\mathbf{I}}$	$\mathcal{D}_{\mathbf{II}}$
Purpose	Training	Test	Test	Test
Original dataset	$\mathcal{S}_{tr}$	$\mathcal{S}_{ts}$	$\mathcal{S}_{ts}$	$\mathcal{S}_{ts}$
No. images	18000 per k (72000 total)	2000 per k (8000 total)	100 per each (k, $\{\text{QF}_{1,i}\}$, h × w)	100 per each (k, $\{\text{QF}_{1,i}\}$, h × w)
DJPEG	Type I and II (50% each)	Type I and II (50% each)	Type I	Type II
Setting	$\{\text{QF}_{1,i}\}$, h×w randomly chosen	$\{\text{QF}_{1,i}\}$, h×w randomly chosen	h×w = $\{96 \times 96$, $128 \times 128\}$	h×w = $\{96 \times 96$ $128 \times 128\}$

REFERENCES

[1] D. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "RAISE: A Raw Images Dataset for Digital Image Forensics", *Proc. the 6th ACM Multimedia Systems Conference*, 2015.

[2] Y. Niu, B. Tondi, Y. Zhao and M. Barni,"Primary quantization matrix estimation of Double compressed JPEG Images via CNN", *IEEE Signal Process. Lett.*, 2020.