# Artificial Intensity Remapping: Learning Multimodal Image Descriptors without Multimodal Image Data

**Youngwook Paul Kwon**
University of California at Berkeley
Berkeley, CA 94720
young@berkeley.edu

**Sara McMains**
University of California at Berkeley
Berkeley, CA 94720
mcmains@berkeley.edu

## Abstract

We present a novel, simple, and aggressive data augmentation strategy called Artificial Intensity Remapping (AIR). AIR artificially generates challenging but realistic training input to improve generalization of learning, making a trained model more effective even for adversarial input. In building a deep-learning-based image descriptor model, experiments show that our model trained using only non-multimodal data with AIR outperforms state-of-the-art algorithms not only for non-multimodal data but also for multimodal data. This multimodal testing demonstrates the benefits of AIR in addressing the fact that successful performance in deep learning, with its huge number of parameters, may result from overfitting to a given dataset.

## 1 Introduction

Finding correspondences is a fundamental task in computer vision. A key step is to assign discriminative descriptors to local features, such that descriptors are distinctive for each feature within an image, and at the same time, consistent between corresponding features across images.

**Hand-crafted vs. Data-driven Descriptors** Most descriptors are hand-crafted [7, 10], relying on gradient histograms, intensity distributions, etc. More recently, approaches exploiting deep learning to generate descriptors show impressive performance [9, 11, 5], typically using a Siamese network [1] for generating a descriptor embedding (mapping).

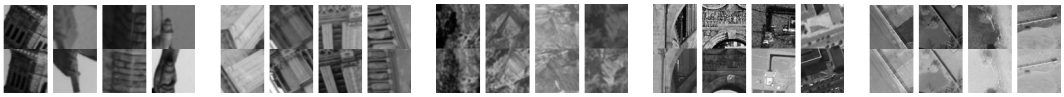
**Multimodal Input** Whether hand-crafted or data-driven, existing methods generally do not exhibit a similar level of performance under severe appearance changes (e.g., day vs. night, different sensors, years apart). We denote these kinds of image pairs *multimodal* in a broad sense of the word (see some examples in Fig. 1). Relatively less research has been devoted to such input, with some recent successes reported by [3, 6], based on hand-crafted descriptors relying on symmetric appearance [3] or line segment distribution [6]. However, they may fail for inputs where their assumptions do not apply (i.e., lack of symmetric or linear features).

**Data augmentation** Since deep neural networks need to be trained on a huge number of training images to achieve satisfactory performance, data augmentation can boost performance. Multiple combinations of horizontal/vertical flipping, cropping, color jittering, etc. are popularly used. Krizhevsky et al. [4] proposed using PCA to guide altering RGB channel values in training images, and reduced the top-1 error rate by over 1% in the ImageNet 2012 competition.

We propose deep-learning-based descriptors that are effective for multimodal input. Our research is novel in that we achieve this not by learning from a target multimodal dataset but from a non-

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Figure 1: Examples of multimodal image pairs from [3]



(a) LY [2] (b) ND [2] (c) YO [2] (d) MU1 [3] (e) MU2 [8]

Figure 2: Examples of different datasets: Statue of Liberty (LY), Notre Dame (ND), Half Dome in Yosemite (YO) are from [2]. Multimodal sets, MU1 and MU2, are cropped from [3] and [8].

multimodal dataset. To do so, we need to ensure that the learning process does not overfit so that it will be robust to extremely diverse input.

In many cases in machine learning research, evaluation of an algorithm is conducted by separating a single dataset into training and test sets. Despite this separation, the sets may reflect undesired and unintended correlation (e.g., a certain style of a sensor), which will lead to the trained models being less robust for images in the wild.

For better generalization, we introduce a new data augmentation strategy, *Artificial Intensity Remapping* (AIR). We aggressively generate more challenging but realistic input for training. To demonstrate the efficacy of AIR, we train a model in a manner similar to [9], and evaluate with multimodal images, which would normally be considered to be adversarial input. Our experiments shows that a model powered by AIR outperforms not only for the intra-data test but also for the cross-data test. Note that this is also beneficial considering the limited availability of multimodal datasets.

## 2 Datasets

For data-driven local image descriptor learning, Brown et al. [2] provide the Multi-view Stereo Correspondence (MVS) dataset of 64×64 gray-scale image patches sampled from the Statue of Liberty (LY), Notre Dame (ND) and Yosemite (YO), with labeled correspondence relationships. Some corresponding patches are shown in Fig. 2a–2c. Following Simo-Serra et al. [9], for training, we select 100,000 positive samples (corresponding patch pairs) and negative samples (non-corresponding random patch pairs) for each subset of the MVS dataset (LY, ND, YO). We train a model with the sampled patches of two subsets (e.g., ND and YO), and test using the other subset (e.g., LY). The testing process is discussed in Section 4. Despite training and test set separation, the three subsets may share a certain style of appearance. Overfitting may occur during learning, and if so, the resulting descriptors may not work as well for input from different datasets.

We create an adversarial test set by extracting multimodal patches as shown in Fig. 2d–2e, from [3, 8]. In contrast to the image patches of the MVS dataset, these multimodal datasets with various modalities [3] or electro-optic (EO) vs. infra-red (IR) [8] consist of corresponding full image pairs with ground truth transformations between pairs. For a given image pair, we register the images using the transformation, and crop $64 \times 64$ image patches located at every stride of 32 pixels. From [3] (46 image pairs) and [8] (980 image pairs), we create around 15,000 (MU1) and 100,000 (MU2) corresponding patch pairs, respectively. [1] Note that we do not train any models from these multimodal datasets (MU1 and MU2); for multimodal testing, we train a model with all three MVS subsets (LY, ND, YO).

## 3 Artificial Intensity Remapping

A more robust system should work on input beyond the style(s) available in the training set. What if we expose the machine to various versions of corresponding patches? Under this reasoning, we generate randomized *artificial intensity remappings*. One can think of a mapping function $F : x \rightarrow y$ $(0 \leq x, y \leq 255)$ that changes the look of a patch, i.e., changes intensity from $x$ to $y$.

[1] Since most image descriptors consider gray-scale input, all color images in [3] are converted to gray-scale.

(a) Randomly generated mapping functions (b) One original patch and its ten derived patches

Figure 3: Examples of Artificial Intensity Remapping

We first define $k$ control points, $(x_i, z_i)$ for $i = 1, 2, \ldots, k$:

$$x_i \leftarrow 255(i-1)/(k-1) \quad (k \text{ uniform subdivisions within } [0, 255]) \tag{1}$$

$$r_i \leftarrow \text{uniform random sample within } [0, 1) \tag{2}$$

$$a \leftarrow \text{random binary sample, either 0 or 1} \quad (\text{determines whether to reverse gradient}) \tag{3}$$

$$s_i \leftarrow (-1)^a \sum_{1}^{i} r_i \quad (\text{accumulated sum with sign}) \tag{4}$$

$$z_i \leftarrow 255/(\max_i(s_i) - \min_i(s_i))(s_i - \min_i(s_i)) \quad (\text{range normalization}) \tag{5}$$

The $s_i$s are the accumulated sum of random samples ($r_i$s), and $a$ determines whether they are increasing positive values or decreasing negative values. The $z_i$s are the normalization of the $s_i$s such that the minimum and maximum values are 0 and 255. Taking $(x_i, z_i)$ as $k$ control points, equally spaced on the $x$ axis, we interpolate them with a second order interpolating spline (i.e., going through each control point with continuous first derivatives). This spline defines a mapping function $Z$ that maps a pixel intensity $x$ to $Z(x)$.

To introduce noise, we also perturb the function $Z$ at every possible integer point $x_j$ by a random amount less than a constant perturbation parameter, $p$.

$$x_j \leftarrow j \quad (j = 0, 1, \ldots, 255) \tag{6}$$

$$b_j \leftarrow \text{uniform random sample within } [-1, 1] \tag{7}$$

$$y_j \leftarrow Z(x_j) + b_j \cdot p \quad (\text{perturbation}) \tag{8}$$

Then we re-interpolate these new $(x_i, y_i)$ points using linear interpolation to obtain our final function $F$ that maps a pixel intensity $x$ to $F(x)$.

We show two example random mapping functions and patches before and after applying the functions in Fig. 3a, and visualize ten variations of resulting patches in Fig. 3b ($k = 7, p = 10$). When training, we generate a different mapping function for every patch and apply it to the patch, followed by random horizontal and/or vertical flipping and image whitening.

## 4 Experiments

**Setup** To show the efficacy of AIR, we set up our network similarly to [9, 5], which show state-of-the-art performance. The architectural differences are summarized in Table. 1. As a key strategy, [9] introduces a mining scheme whereas [5] introduces an extra global term in their loss function. We also adopt the mining scheme of [9].

**Evaluation** We follow the evaluation method of [9]. The testing process is as follows. We randomly pick 8,000 patches. For each of the selected patches, we select one corresponding patch and 1,000 other random patches, and calculate the descriptor distances of 1 true pair and 1,000 false pairs. PR (precision-recall) curves are calculated from the descriptor distances.

The PR curves of the cross-subset MVS tests (LY, ND, YO) are shown in Fig. 4(a–c), and AUCs (area under curve) are shown in Table. 2a [2] [3] [4]. Numbers and letters in parentheses represent descriptor dimension and type (f:floating point, b:binary), respectively. The model trained with AIR had the best overall performance. For the multimodal tests (MU1, MU2), we compare AIR to recently reported deep-learning-based algorithms as well as SIFT. As seen in the PR curves (Fig. 4 d–e) and AUCs (Table. 2b), our model powered by AIR presents the best performance for the multimodal input.

[2] In performance reporting, we reconstructed the MVS training and test sets following the randomized procedure of [9]; therefore our test sets are not exactly the same as in [9].

[3] Because Global [5] provides only LY-trained, ND-trained, and YO-trained models, we record whichever gives the best performance in Table. 2a and 2b.

[4] Because SIFT is a hand-crafted descriptor, the "training" columns in Table. 2a and 2b are not applicable.

| | architecture | dimension | connectivity | pooling | activation |
|---|---|---|---|---|---|
| [9] | C(7,2,32)-P(2)-C(6,3,64)-P(3)-C(5,4,128) | 128 | sparse | $L_2$ | tanh |
| [5] | B(7,3,96)-P(2)-B(5,1,192)-P(2)-B(3,1,256)-B(1,1,256)-C(1,1,1) | 256 | regular | max | ReLU |
| AIR | C(7,3,96)-P(2)-C(5,1,192)-P(2)-C(3,1,256)-C(1,1,256) | 256 | regular | max | ReLU |

Table 1: Architectural comparison with the most closely related work: $\mathrm{C}(w, s, n)$ denotes a convolution layer with $n$ filters of size $w \times w$ and stride $s$; $\mathrm{B}(w, s, n)$ denotes $\mathrm{C}(w, s, n)$ with batch normalization; $\mathrm{P}(w)$ denotes a pooling layer of size $w \times w$ with stride $w$.

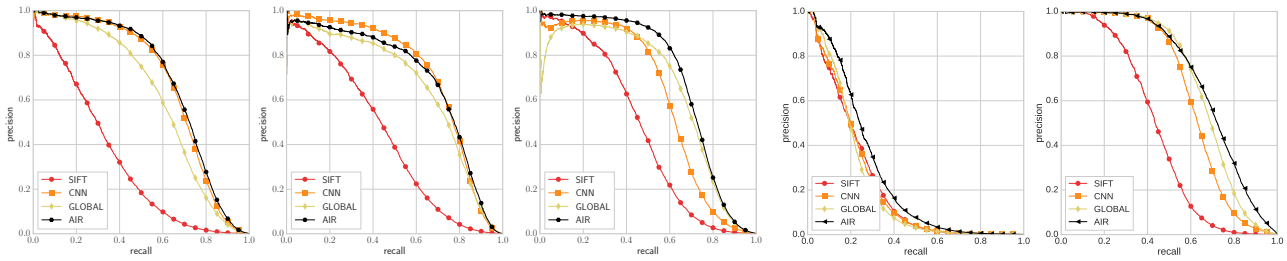
| training | test | SIFT (128f) | CNN (128f) | Global (256f) | AIR (256f) |
|---|---|---|---|---|---|
| ND,YO | LY | 0.314 | 0.678 | 0.612 | **0.690** |
| LY,YO | ND | 0.422 | **0.715** | 0.661 | 0.697 |
| LY,ND | YO | 0.451 | 0.604 | 0.651 | **0.703** |
| average | | 0.396 | 0.666 | 0.641 | **0.697** |

(a) Intra-dataset cross tests

| training | test | SIFT (128f) | CNN (128f) | Global (256f) | AIR (256f) |
|---|---|---|---|---|---|
| LY,ND,YO | MU1 | 0.216 | 0.217 | 0.211 | **0.264** |
| LY,ND,YO | MU2 | 0.437 | 0.633 | 0.681 | **0.708** |
| average | | 0.327 | 0.425 | 0.446 | **0.486** |

(b) Inter-dataset cross tests

Table 2: AUC of precision-recall curves for intra-dataset cross tests (cross-subset MVS tests) and inter-dataset cross tests (multimodal tests) using SIFT[7], CNN[9], Global[5] and AIR




(a) LY (b) ND (c) YO (d) MU1 (e) MU2

Figure 4: Precision-recall curves for each test case

## 5 Conclusion

We present a novel data augmentation scheme, Artificial Intensity Remapping (AIR). AIR artificially generates adversarial but still realistic input to help learning for better generalization. Overall, our model trained with AIR using only non-multimodal training data outperforms existing methods on various tests, including multimodal data.

## References


[1] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Säckinger, and R. Shah. Signature verification using a "Siamese" time delay neural network. *IJPRAI*, 07, 1993.

[2] M. Brown, G. Hua, and S. Winder. Discriminative learning of local image descriptors. *TPAMI*, 33, 2011.

[3] D. C. Hauagge and N. Snavely. Image matching using local symmetry features. *CVPR*, 2012.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

[5] B. G. V. Kumar, G. Carneiro, and I. Reid. Learning local image descriptors with deep Siamese and Triplet convolutional networks by minimising global loss functions. *CVPR*, 2016.

[6] Y. P. Kwon, H. Kim, G. Konjevod, and S. McMains. DUDE (DUality DEscriptor): A robust descriptor for disparate images using line segment duality. *ICIP*, 2016.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[8] S. Razakarivony and F. Jurie. Vehicle detection in aerial imagery : A small target detection benchmark. *JVCIR*, 34, 2015.

[9] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. *ICCV*, 2015.

[10] Z. Wang, B. Fan, and F. Wu. Local intensity order pattern for feature description. *ICCV*, 2011.

[11] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. *CVPR*, 2015.