

Joint detection and matching of feature points in multimodal images

Anonymous CVPR submission

Paper ID ****

Abstract

In this work, we propose a novel Convolutional Neural Network (CNN) architecture for the joint detection and matching of feature points in images acquired by different sensors using a single forward pass. The resulting feature detector is tightly coupled with the feature descriptor, in contrast to classical approaches (SIFT, etc.), where the detection phase precedes and differs from computing the descriptor. Our approach utilizes two CNN subnetworks, the first being a Siamese CNN and the second, consisting of dual non-weight-sharing CNNs. This allows simultaneous processing and fusion of the joint and disjoint cues in the multimodal image patches. The proposed approach is experimentally shown to outperform contemporary state-of-the-art schemes when applied to multiple datasets of multimodal images by reducing the matching errors by 50%-70% compared with previous works. It is also shown to provide repeatable feature points detections across multi-sensor images, outperforming state-of-the-art detectors such as SIFT and ORB. To the best of our knowledge, it is the first unified approach for the detection and matching of such images.

1. Introduction

The detection and matching of feature points in images is a fundamental task in computer vision and image processing that is applied in common computer vision tasks such as image registration [41], dense image matching, [37] and 3D reconstruction [1], to name a few. The term *feature point* relates to the center of an image patch, that is expected to be salient and repeatedly detected in multiple images of the same scene, which might differ by pose and appearance [25]. A *detector* identifies the spatial location of a feature point, and the surrounding patch is encoded by a *descriptor*.

The detection and matching of feature points in multimodal images, as depicted in Fig. 1, is of particular interest in remote sensing [17, 23, 19, 15] and medical imaging [35], as the fusion of such images, provides information synergy. The acquisition of the same scenes by different sensors might result in significant appearance variations,

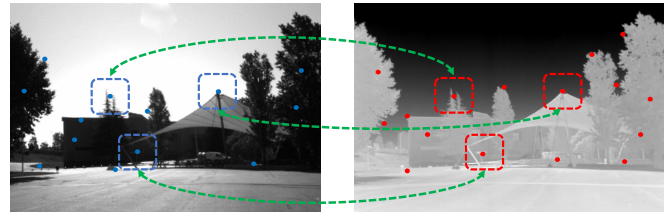


Figure 1. The multisensor patch matching problem. The matched optical (left) and IR (right) images differ by significant appearance changes due to the dissimilar physical characteristics captured by the different sensors. The images are part of the LWIR-RGB dataset [2]. The feature points in both images were detected using the proposed scheme.

that are often nonlinear and unknown apriori, such as non-monotonic intensity mappings, contrast reversal, and non-corresponding edges and textures.

The registration of the multimodal input images \mathbf{I}_1 and \mathbf{I}_2 can be formulated as the estimation of a parametric (rigid, affine, etc.) global transformation T , by minimizing an appearance invariant similarity measures $\phi(\mathbf{I}_1, \mathbf{I}_2)$, such as mutual information [38]

$$T^* = \arg \min_T \phi(\mathbf{I}_1, T(\mathbf{I}_2)). \quad (1)$$

$T(\mathbf{I}_2)$ is the image \mathbf{I}_2 warped towards \mathbf{I}_1 , according to T . Gradient-based approaches were applied by Irani et al. [17], and Keller et al. [19] to appearance-invariant image representations to solve Eq. 1 iteratively.

Other multimodal registration schemes are based on matching local image features such as patches, contours [23], and corners. Such approaches match the sets of interest points $S_1 \in \mathbf{I}_1$ and $S_2 \in \mathbf{I}_2$, where each feature point is first detected and then encoded by a robust appearance-invariant descriptor. A pair of descriptors can be matched by computing their L_2 distance.

Such descriptors were commonly derived by extending unimodal descriptors such as SIFT [25] and Daisy [36] to the multimodal case [10, 16, 2, 5, 21].

Convolutional Neural Networks (CNNs) were applied to feature point matching [3], by training data-driven multimodal image descriptors. These CNNs are trained by opti-

mizing a Hinge Loss applied to an L_1 or L_2 metrics, while others [40, 4, 12, 30] aim to compute a similarity score between image patches by optimizing the Cross-Entropy loss by classifying the pairs of patches as same/not-same. Such approaches utilize Siamese CNNs [3] consisting of weight sharing sub-networks.

The upside of L_2 -based representations compared to those computed using the Cross-Entropy loss is their reduced computational complexity when applied to match sets of feature points detected in a pair or set of images. As an image typically contains $n = 300 \div 500$ feature points, matching a pair of images requires $n^2/2$ point-to-point similarity evaluations. K nearest neighbors (KNN) similarity search via L_2 -based representations can be computationally accelerated using metric embedding schemes such as Locality Sensitive Hashing (LSH) [14] and MinHash [8].

Feature detectors are commonly applied to each image separately, without relating the detections in one modality to the other. A detector aims to detect the spatial location of the feature point and estimate its local scale and orientation. The location is often determined using a corner detector [25], and the local scale is computed using the Difference of Gaussians (DoG) operator and its approximations [24, 25].

In this work, we propose a CNN-based metric learning approach for the joint detection and matching of feature points in multimodal images, using a single forward pass. When applied to image patches, the proposed scheme computes the corresponding descriptors and similarity. But, when applied to full-scale images, the proposed fully convolutional CNN computes a grid of descriptors. By propagating back through the corresponding CNN activations and layers, the locations of feature points corresponding to each descriptor are detected. As the CNN is trained to optimize the descriptors' matching, the proposed detector is tightly coupled with the feature descriptor, in contrast to classical approaches such as SIFT [25] and its (many) extensions, where the detector is computed separately before computing the descriptor. To the best of our knowledge, we introduce the first unified approach for the detection and matching of multi-modality images.

In particular, we present the Hybrid CNN architecture consisting of both a Siamese sub-network and a dual-channel non-weight-sharing asymmetric sub-network. The use of the asymmetric sub-network is due to the inherent asymmetry in the multisensor matching problem, where the heterogeneous inputs might differ significantly, and thus require different processing implemented by the asymmetric sub-network. In particular, each branch of the asymmetric sub-network estimates a *modality-specific* adaptive representation of the multisensor patches.

Thus, we aim to leverage both the joint and disjoint attributes in the multimodal images, using the Siamese and

Asymmetric subnets, respectively. Siamese sub-networks were previously shown [3] to yield accurate matching results, and are outperformed by the proposed Hybrid scheme. The Siamese and Asymmetric subnets are trained by corresponding losses, and their outputs are merged and optimized to yield a fused image representation.

In particular, we propose the following contributions:

First, we present a novel approach for the joint detection and matching of feature points in multi-modality images.

Second, the proposed scheme is implemented using a novel Hybrid CNN architecture consisting of both a Siamese and asymmetric sub-networks, able to leverage both the joint and disjoint cues in multimodal patches, to determine their similarity.

Third, we show that training the proposed Hybrid CNN by multi-task learning improves the descriptors' matching accuracy.

Last, the proposed scheme was experimentally shown to outperform contemporary approaches when applied to state-of-the-art multimodal image matching benchmarks [3, 4, 12] and feature points detection schemes. The corresponding source code was made publicly available¹.

2. Related work

Common approaches for computing appearance-invariant image representations of multisensor images utilize salient image edges and contours. Irani et al. [17] suggested a coarse-to-fine scheme for estimating the global parametric motion (affine, rigid) between multimodal images, using the magnitudes of directional derivatives as a robust image representation. The correlation between these representations is maximized using iterative gradient methods and a coarse-to-fine formulation.

The "Implicit Similarity" formulation by Keller et al. [19] is an iterative scheme utilizing gradient information for global alignment. A set of pixels with maximal gradient magnitude is detected in one of the input images, rather than contours and edges as in [17]. The gradient of the corresponding points in the second image is maximized with respect to a global parametric motion, without explicitly maximizing a similarity measure.

The seminal work of Viola and Wells [38] on applying the mutual information (MI) similarity to multisensor image matching, utilized a statistical representation of the images while optimizing their mutual information with respect to the motion parameters.

Modality-invariant descriptors were often derived by modifying the seminal SIFT descriptor [10, 26]. Contrast-invariance was achieved by mapping the gradient orientations of the interest points from $[0, 2\pi)$ to $[0, \pi)$. Hasan et al. showed that such descriptors mitigate the matching accu-

¹<https://github.com/eladbb/HybridSiamese>

racy [16], and further modified the SIFT descriptor [15] by thresholding gradient values to reduce the effect of strong edges. An enlarged spatial window with additional sub-windows was used to improve the spatial resolution.

The Self Similarity Matching (SSM) approach by Shechtman et al. [33] is a geometric descriptor encoding the local structure of a feature point in an image, by correlating a central patch to all adjacent patches within a predefined radius. The Dense Adaptive Self-Correlation (DASC) descriptor, by Kim et al. [20], extended the SSM approach, by computing the self-similarity measure as an adaptive self-correlation between randomly sampled patches.

Aguilera et al. [2] used a histogram of contours and edges instead of a histogram of gradients to avoid the ambiguity of the SIFT descriptor when applied to multimodal images, while the dominant orientation was determined similarly. This approach was extended by the same authors by using multi-oriented and multi-scale Log-Gabor filters [5]. The Duality Descriptor (DUDE) multimodal descriptor was proposed by Kwon et al. [21], where each line segment near a keypoint is encoded by a 3D histogram of radial, angular and length parameters. This approach encodes the geometry of the line segment and is invariant to appearance variations.

With the emergence of CNNs as the state-of-the-art approach to a gamut of computer vision problems, CNNs were applied to patch matching. Zagoruyko and Komodakis [40] proposed several CNN architectures, such as a Siamese CNN with an L_2 or Cross-Entropy losses, for matching single modality patches, and a CNN, where the input patches are stacked as different image channels. Aguilera et al. [3] applied the approaches by Zagoruyko and Komodakis [40] to matching multimodal patches and showed that the resulting CNN outperformed the state-of-the-art multimodal descriptors.

To alleviate the computational complexity of the stacked approach when applied to sets of feature points, Aguilera et al. proposed the Q-Net CNN [4] that was trained using an L_2 loss. The Q-Net CNN consists of four weight sharing sub-networks and two corresponding pairs of input patches, that allow hard negative mining [28]. This approach was shown to achieve state-of-the-art accuracy when applied to the Vis-Nir benchmark [3].

Recent work by En et al. [12] introduced a hybrid Siamese CNN, similar to the proposed scheme, denoted as TS-Net, for multimodal patch matching, consisting of Siamese and asymmetric sub-networks, utilizing a Cross-Entropy loss. Contrary to the proposed scheme, this approach does not compute L_2 -optimized patch encodings that are essential for the matching of images, typically consisting of 300-500 feature points. Each of the sub-networks outputs a scalar Softmax prediction, and the different predictions are merged using a Fully Connected (FC) layer.

The proposed scheme is experimentally shown in Section 5 to outperform the TS-Net [12] results significantly.

Metric learning was applied by Quan et al. [29] to learn the shared feature space of multi-spectral patches, by progressively comparing spatially connected features, using a discrimination constraint (SCFDM). This approach was extended by the same authors, by deriving the AFD-Net [30], that learns multiscale joint multi-spectral features using a CNN consisting of two subnetworks. The activations maps at different layers are subtracted, and the differences are propagated through multiple FC layers. Thus, this approach does not compute a descriptor, and the matching of two images having n feature points each, entails $n^2/2$ forward passes of the CNN, in contrast, to the single forward pass required by the proposed scheme.

The joint detection and matching of feature points in *single* modality images were studied by Dusmanu et al. [11] using a Siamese network, where the descriptors are trained using a triplet ranking loss, and the features are detected as the local maxima of the last activation map. The corresponding CNN was implemented without pooling layers to relate the detections in the last activation map to the source (finest) image resolution. The model was trained using pixel correspondences computed by large-scale SfM reconstructions. An image pyramid consisting of three resolutions is used to account for scale variations, by computing the descriptors in all scales.

Simeoni et al. [34] proposed a multi-scale feature detection scheme for *single* modality images by detecting local maxima in the activation maps over multiple activation layers. The activations are localized per channel using the maximally stable extremal regions (MSER) blob detector. As in [11], a Siamese network is used to match the corresponding detections in the training images.

3. Detection and matching of multi-modal feature points

In this section, we present the Hybrid multisensor detection and matching scheme depicted in Fig. 2. Let $\mathbf{x}_i \in \mathbf{I}_1$ and $\mathbf{y}_i \in \mathbf{I}_2$ be a pair of multi-dimensional image patches acquired by different modalities. We aim to compute a corresponding representation, $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{y}}_i$, respectively.

The Hybrid network learns both the joint and disjoint characteristics of the multisensor patches, using both Siamese (symmetric) and asymmetric (non-weight-sharing) networks. The Siamese network learns *the same* mapping for both input modalities, denoted as $W_s(\mathbf{x}_i)$ and $W_s(\mathbf{y}_i)$ in Fig. 2, respectively, allowing to encode the *same* characteristics between the images. The asymmetric network estimates *different*, modality-specific representations, $W_x(\mathbf{x}_i)$ and $W_y(\mathbf{y}_i)$, for each modality, respectively. The outputs of the symmetric and asymmetric sub-networks are concate-

nated

$$H_x(\mathbf{x}_i) = [W_s(\mathbf{x}_i), W_x(\mathbf{x}_i)]^T, \quad (2)$$

$$H_y(\mathbf{y}_i) = [W_s(\mathbf{y}_i), W_y(\mathbf{y}_i)]^T,$$

and the resulting Hybrid representation is given by

$$\hat{\mathbf{x}}_i = FC_x(H_x(\mathbf{x}_i)), \quad (3)$$

$$\hat{\mathbf{y}}_i = FC_y(H_y(\mathbf{y}_i))$$

where FC_x and FC_y are FC layers.

The proposed CNN is fully convolutional and can thus be applied to images of varying dimensions. When applied to 64×64 image patches, the Hybrid CNN yields a pair of $\{\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i\} \in \mathbb{R}^{128}$ descriptors. But, when applied to larger images, an activation map $\{\hat{\mathbf{x}}_{i,j}, \hat{\mathbf{y}}_{i,j}\} \in \mathbb{R}^{128}$ is computed. Each descriptor relates to a particular patch in the image, according to its footprint. We show that by backtracking through the CNN activations, down to the input layer, we detect the location of the corresponding feature points in the finest image resolution, as detailed in Section 3.2.

The proposed CNN is trained using multi-task learning. The losses L_S and L_A in Fig. 2 optimize the symmetric and asymmetric subnetworks, respectively. This was shown to improve the matching accuracy, as these losses (L_S and L_A) optimize CNNs having less number of parameters compared to the full Hybrid CNN. The unified Hybrid representation, as in Eq. 3, is trained using the L_H loss.

We used both the Binary Cross-Entropy (BCE) and Hinge Loss applied to the L_2 metric loss, such that $\{L_S, L_A, L_H\}$ are either all BCE or L_2 losses. The choice of the loss relates to the particular task of the multi-modality descriptors $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{y}}_i$. For the BCE-based descriptors, a Softmax layer outputs the matching probability of the input patches, while the L_2 loss yields a descriptor embedded in a Euclidean space. Such descriptors can be utilized in efficient large-scale descriptor retrieval schemes, where K -nearest-neighbors (KNN) L_2 search can be efficiently implemented using LSH [14] and MinHash [8].

3.1. CNN architecture

The proposed networks consist of a Siamese (symmetric) and asymmetric networks, as depicted in Fig. 2. The Siamese network consists of the two weight-sharing networks W_s , while the asymmetric network consists of the non-weight-sharing networks $\{W_X, W_Y\}$ applied to the inputs X and Y , respectively. When the losses $\{L_S, L_A, L_H\}$ are all L_2 Hinge losses, W_X, W_Y and W_s are detailed in Table 1. Similarly, when the losses $\{L_S, L_A, L_H\}$ are all BCE losses, W_X, W_Y and W_s are given by Table 2, and the overall training loss is given by

$$L = L_S + L_A + L_H \quad (4)$$

Layer	Output	Kernel	Stride	Pad
Conv0	$64 \times 64 \times 32$	5×5	1	2
Pooling	$32 \times 32 \times 32$	3×3	2	-
Conv1	$32 \times 32 \times 64$	5×5	1	2
Pooling	$16 \times 16 \times 64$	3×3	2	-
Conv2	$16 \times 16 \times 128$	3×3	1	1
Pooling	$8 \times 8 \times 128$	3×3	2	-
Conv3	$6 \times 6 \times 256$	3×3	1	0
Conv4	$4 \times 4 \times 256$	3×3	1	0
FC	1×128	-	-	-
Unit norm	1×128	-	-	-

Table 1. The CNN architecture of the sub-networks using the L_2 Hinge loss. Each sub-network accepts a 64×64 patch and outputs a 128×1 descriptor.

Layer	Output	Kernel	Stride	Pad
Conv0	$64 \times 64 \times 32$	5×5	1	2
Pooling	$32 \times 32 \times 32$	3×3	2	-
Conv1	$32 \times 32 \times 64$	5×5	1	2
Pooling	$16 \times 16 \times 64$	3×3	2	-
Conv2	$16 \times 16 \times 128$	3×3	1	1
Pooling	$8 \times 8 \times 128$	3×3	2	-
Conv3	$6 \times 6 \times 256$	3×3	1	0
Conv4	$4 \times 4 \times 256$	3×3	1	0
Conv5	$2 \times 2 \times 256$	3×3	1	0
FC	1×128	-	-	-

Table 2. The CNN architecture of the sub-networks using the Cross-Entropy loss. Each sub-network accepts a 64×64 patch and outputs a 128×1 descriptor.

3.2. The detection of feature points in multi-modality images

The proposed scheme consists of fully convolutional CNNs that when applied to an image yield an activation map $\{\hat{\mathbf{x}}_{i,j}, \hat{\mathbf{y}}_{i,j}\} \in \mathbb{R}^{256 \times 1}$. Thus, a feature point $\mathbf{v} = (l, k)$ is a pixel location, such that the feature points $\mathbf{v}_x \in \mathbf{I}_1$ and $\mathbf{v}_y \in \mathbf{I}_2$ relate to the descriptors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, respectively.

The fundamental property of a feature point is its *repeatability* [25, 27], implying that corresponding points $\mathbf{v}_x \in \mathbf{I}_1$ and $\mathbf{v}_y \in \mathbf{I}_2$ relate to *joint* image content in both images. Thus, we propose to detect the feature points by analyzing the *joint* representation encoded by the Siamese subnetworks $\hat{\mathbf{x}}_i^s = W_s(\mathbf{x}_i)$ and $\hat{\mathbf{y}}_i^s = W_s(\mathbf{y}_i)$ (as in Fig. 2) to detect $\mathbf{v}_x \in \mathbf{I}_1$, and $\mathbf{v}_y \in \mathbf{I}_2$, respectively.

Next, we detail the detection of $\mathbf{v}_x \in \mathbf{I}_1$ given $\hat{\mathbf{x}}_i^s$ and W_s , where the same approach mutatis mutandis is applied to detect $\mathbf{v}_y \in \mathbf{I}_2$. We aim to detect the feature point \mathbf{v}_x in the finest image resolution by backtracking the source of the activation $\hat{\mathbf{x}}_i^s$ through W_s , down to the first activation layer.

The W_s CNN, detailed in Tables 1 and 2, consists of padded convolutions and pooling layers. Symmetric (i.e.

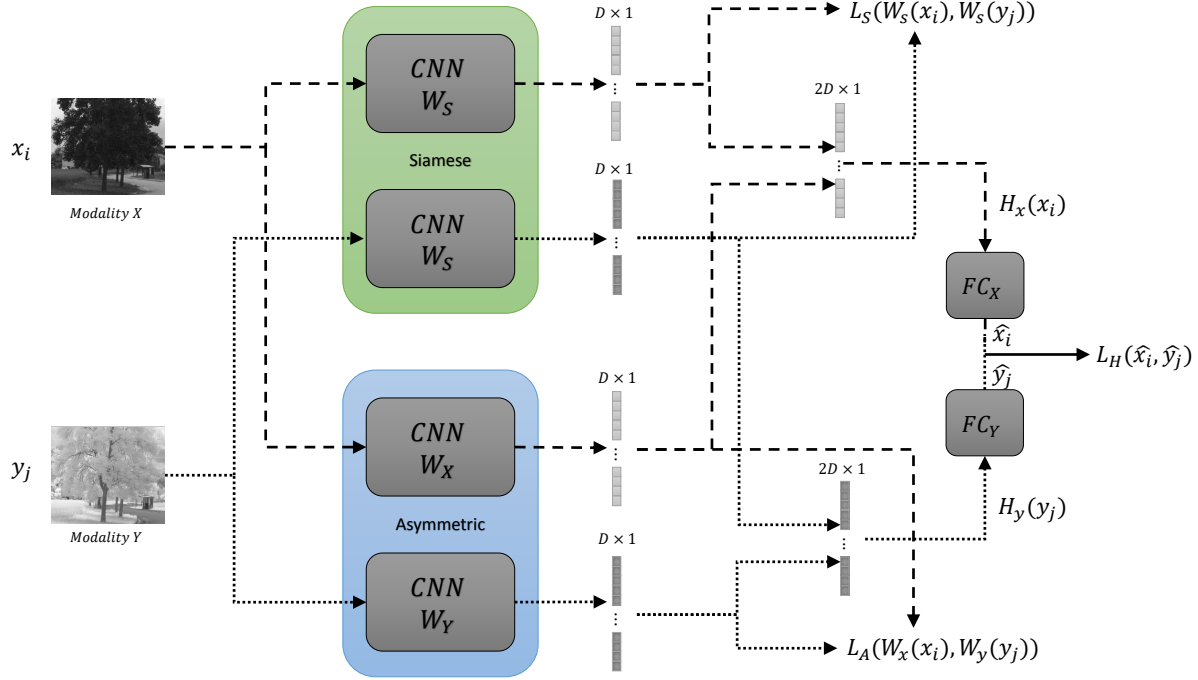


Figure 2. The proposed hybrid matching model, consisting of two sub-networks: a siamese subnetwork and an asymmetric subnetwork with non-shared weights. The siamese branch consists of a pair of W_s CNNs and is trained by the loss L_s . The asymmetric branch consists of the W_x and W_y CNNs and is trained by the loss L_A . The symmetric and asymmetric representations are merged and trained by the loss L_H .

3×3) padded convolutions do not change the locations of the activations, while max-pooling layers propagate the content of a single location in the activation map. Let $\hat{x}^l \in \mathbb{R}^{w_l \times h_l \times c_l}$ be an activation map at level l of spatial dimensions $w_l \times h_l$ and c_l channels. Each element in \hat{x}^l is the result of the max-pooling layer applied to the preceding activation map \hat{x}^{l-1} .

Let $\hat{x}_{in} \in \mathbb{R}^{W \times H \times C}$ and $\hat{x}_{out} \in \mathbb{R}^{\frac{W}{k} \times \frac{H}{k} \times C}$ be the input and output layers, respectively, of a $k \times k$ max-pooling layer. Max-pooling is applied channel-wise, where for each $k \times k \times 1$ tensor in \hat{x}_{in} , the entry having the maximal value is propagated forward. Thus, the entries in a *single* spatial location $\hat{x}_{out}(\frac{w}{k}, \frac{h}{k}, c)$, $c = 1..C$ might relate to *multiple* spatial locations in \hat{x}_{in} . To relate these entries to a *single* spatial location in \hat{x}_{in} , we only backtrack the location of a *single* maximal entry $\hat{x}_{out}(\frac{w}{k}, \frac{h}{k}, c^*)$ such that

$$c^* = \underset{c}{\operatorname{argmax}} \left\{ \hat{x}_{out} \left(\frac{w}{k}, \frac{h}{k}, c \right) \right\}, c = 1..C, . \quad (5)$$

The proposed detection approach differs from Dusmanu et al. [11] that compute local (per layer) detection scores, and merge the scores from different layers, having different spatial resolutions, by bilinearly interpolating the lower-resolution score maps. Simeoni et al. [34] utilize a CNN with no pooling layers to avoid the need for backtracking through the activation layers. Both these schemes are only

applicable to single-modality images.

4. Discussion

The proposed scheme is the first (to the best of our knowledge) to study the joint detection and matching of feature points of multi-modality images, where previous works [12, 3, 4] only dealt with the matching of such images.

Our approach extends previous works by proposing a novel CNN architecture that combines both siamese and asymmetric non-weight sharing CNNs. While asymmetric CNNs have been used in multimodal embedding and classification problems, such as joint image-text embeddings and inference [13, 18], siamese CNNs were shown to provide accurate results [3] when applied to multimodal image matching. We attribute that to edges being a common characteristic of multimodal images, in contrast to image texture that often differs significantly. Some previous works [17, 23, 19, 15] used edges as a robust joint representation for multimodal images. Hence, our approach utilizes additional joint information that can't be captured by applying the same (symmetric) operations to both input modalities.

Contrary to handcrafted approaches in feature detection [25, 7, 27], the proposed scheme does not detect predefined image attributes such as corners or blobs, and we do not apply a dominant-scale or dominant-rotation estimation schemes as in the SIFT detector [25] and its (many) exten-

sions. Thus, the rotation and scale invariance are learned via data augmentation.

The detection and encoding of the feature points are tightly coupled, implying that the detections are trained for optimal matching. In contrast to handcrafted approaches [25, 7, 32] where the feature point is first detected using local image attributes, and a *different* approach is used to compute the descriptor.

5. Experimental Results

The proposed Hybrid scheme was experimentally verified by applying it to multi-spectral image datasets and benchmarks used in contemporary state-of-the-art matching schemes. The first was suggested by Aguilera et al. [3] consisting of a set of matching and non-matching pairs of patches, extracted from nine categories of the public VIS-NIR scene dataset [9]. The feature points were detected by an interest point detector and matched manually. We also used the Vehicle Detection in Aerial Imagery (VEDAI) [31] dataset of multispectral aerial images and the CUHK [39] dataset consisting of 188 faces and corresponding sketches drawn by artists. These multimodal datasets are spatially pre-aligned, same as the VIS-NIR dataset, and were used by En et al. [12] to create annotated training and test sets by extracting corresponding pairs of patches on a uniform grid. We evaluated the matching accuracy of the proposed scheme following the experimental setups and datasets used by Aguilera et al. [3, 4] and En et al. [12] and the results are detailed in Sections 5.1 and 5.2, respectively.

The Hybrid CNN was trained using stochastic gradient descent with a momentum of 0.9, batch size of 128, learning rate of 0.01 and weight decay of 0.0005, where the same hyperparameters were used for training both the L_2 and Cross-Entropy losses. The Hybrid model was trained for 40 and 100 epochs, for the L_2 and Softmax losses, respectively. The networks' parameters were initialized by a normal distribution, where the asymmetric subnets were initialized identically to improve convergence.

In both setups, patches of 64x64 pixels were cropped and augmented by horizontal and vertical flipping, and the patches of each imaging modality were normalized separately by subtracting their mean. Hard negative mining was applied following the Hardnet approach of Mishchuk et al. [28]. The matching accuracy is quantified by the false positive rate at 95% recall (FPR95), same as in [3, 4]. The source code of the proposed scheme was made publicly available².

We compare the results of the proposed Hybrid scheme using both L_2 and Cross-Entropy losses to contemporary state-of-the-art approaches in Sections 5.1 and 5.2. The detection accuracy and an ablation study are discussed in Sec-

tions 5.3 and 5.4.

5.1. VIS-NIR benchmark

The proposed Hybrid scheme was experimentally evaluated using the VIS-NIR dataset [3], and was compared to the state-of-the-art results of Aguilera et al. [3, 4] and En et al. [12] using the same experimental setup. As Aguilera et al. [3, 4] and SCFDM [29] utilized the dataset and experimental setup, we quote their results, while we evaluated En et al. [12] by training their publicly available code³. All of the schemes were trained using the 'Country' category, where we utilized 80% of the given training pairs of patches for training the Hybrid approach using the Cross-Entropy loss, and L_2 Hinge loss. We also compared to state-of-the-art handcrafted multimodal descriptors: LSS [33], DASC [20], LGHD [5], SIFT [25] and MI-SIFT [26] using their publicly available code⁴.

The results are reported in Table 3, where it follows that the LSS and DASC performed worse, as such descriptors aim to encode corresponding large-scale geometrical patterns that might not exist in multisensor images. The proposed Hybrid scheme outperformed the previous methods by showing an average error of 2.52% compared to a mean error of 3.48% of the previously leading approach [29] while outperforming it in **seven out of eight** categories.

5.2. En et al. [12] benchmark

We also evaluated the proposed scheme using the experimental setup proposed by En et al. [12] where the VEDAI [31], CUHK [39] and VIS-NIR [3] datasets were sampled on a uniform grid, and the results are reported in Table 4. We quote the results reported by En et al. [12] for these datasets and setup, and trained the publicly available code⁵ of Aguilera et al. [3, 4] and the proposed Hybrid scheme, using 70% of the data in each dataset for training, 10% for validation and 20% for testing. As before we compared with the results of the SIFT [25] and modality-invariant descriptors: LSS [33], DASC [20], LGHD [5] and MI-SIFT [26].

It follows that the proposed approach significantly outperformed the previous schemes for the VIS-NIR and CUHK datasets yielding an average error that is **threefold more accurate**. For the VEDAI dataset, both Aguilera et al. [3], and the proposed scheme achieved a zero error. This superior performance is achieved without applying hard mining, emphasizing that the Hybrid CNN formulation is the one doing the heavy lifting.

³<https://github.com/ensv/TS-Net>

⁴<https://github.com/ngunsu/LGHD>

<http://www.robots.ox.ac.uk/vgg/software/SelfSimilarity>

<https://github.com/seungrong/DASC>

⁵<https://github.com/ngunsu/lcsis>

<https://github.com/ngunsu/qnet>

²<https://github.com/eladbb/HybridSiamese>

Network/descriptor	Field	Forest	Indoor	Mountain	Old building	Street	Urban	Water	Mean
Engineered Features									
SIFT [25]	39.44	11.39	10.13	28.63	19.69	31.14	10.85	40.33	23.95
Inv SIFT [26]	34.01	22.75	12.77	22.05	15.99	25.24	17.44	32.33	24.42
LGHD [5]	16.52	3.78	7.91	10.66	7.91	6.55	7.21	12.76	9.16
LSS [33]	46	42.48.38	37.14	42.5	42.35	44.5	34.9	46	42.65
DASC [20]	46.68.44	35.38	23.19	41.29	38.07	39.02	12.28	45.6	36.68
Learning-based									
Siamese [3]	15.79	10.76	11.6	11.15	5.27	7.51	4.6	10.21	9.61
Pseudo Siamese [3]	17.01	9.82	11.17	11.86	6.75	8.25	5.65	12.04	10.32
2Channel l[3]	9.96	0.12	4.4	8.89	2.3	2.18	1.58	6.4	4.47
Q-Net 2P-4N [4]	26.03	5.0	9.46	18.21	7.75	11.16	5.46	17.8	12.60
TS-Net [12]	25.45	31.44	33.96	21.46	22.82	21.09	21.9	21.02	24.89
SCFDM [29]	7.91	0.87	3.93	5.07	2.27	2.22	0.85	4.75	3.48
Hybrid CNN									
Hybrid-CE	10.12	6.4	9.34	7.82	4.31	5.01	3.11	7.09	6.74
Hybrid-CE-HM	5.88	1.45	6.93	3.5	2.25	2.37	0.99	3.06	2.97
Hybrid-L_2	19.95	19.49	18.98	18.79	13.99	14.26	14.5	17.46	17.7
Hybrid-L_2-HM	5.62	0.53	3.58	3.51	2.23	1.82	1.90	3.05	2.52

Table 3. Patch matching results evaluated using the VIS-NIR dataset and the patches extracted as in Aguilera et al. [3, 4]. The accuracy is given in terms of the FPR95 score. The schemes names in bold are variations of the proposed scheme, while HM and CE relate to using hard mining and a Cross-Entropy loss, respectively.

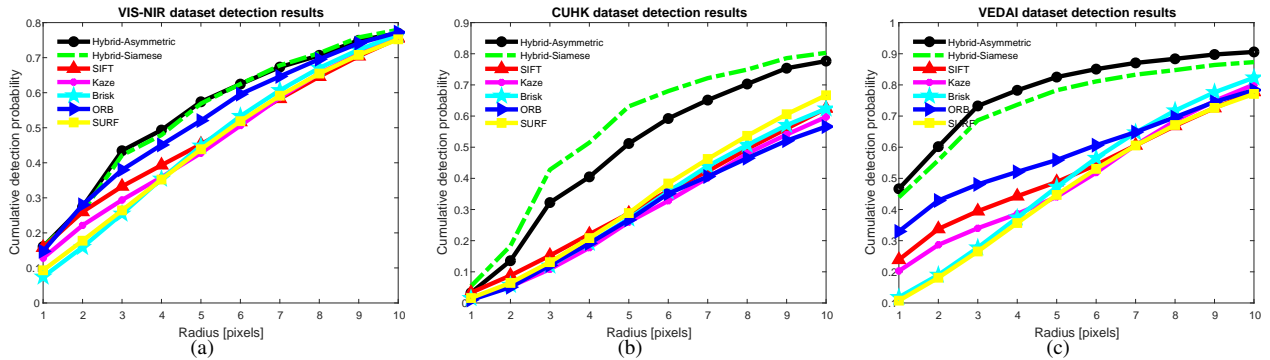


Figure 3. Feature detection results. We report the cumulative detection probability and compare with the SIFT detector. (a) VIS-NIR dataset results. The images and detection maps are 680×1024 and 78×121 , respectively. (b) CUHK dataset results. The images and detection maps are 250×200 and 24×18 , respectively. (c) VEDAI dataset results. The images and detection maps are 512×512 and 57×57 , respectively.

5.3. Features detection

The proposed feature point detection scheme, introduced in Section 3.2, was experimentally evaluated following Mikolajczyk and Schmid [27]. The detection repeatability was measured using the VEDAI [31], CUHK [39] and VIS-NIR [3] image sets consisting of aligned multimodal images. Thus, for each point $\mathbf{v}_x \in \mathbf{I}_1$, there is a corresponding *reference* point $\mathbf{v}_y \in \mathbf{I}_2$. Let $\mathbf{v}_x^d \in \mathbf{I}_1$ be a *detected* feature point, whose corresponding *reference* point is $\mathbf{v}_y \in \mathbf{I}_2$, and let $\mathbf{v}_y^d \in \mathbf{I}_2$ be the *detected* point that is the closest to \mathbf{v}_y . We report the cumulative detection probability

$$P_C = P\left(\|\mathbf{v}_y - \mathbf{v}_y^d\|_2 \leq r\right), \quad r \in \mathbb{R}^+, \quad (6)$$

that is the average probability (over all detected points $\mathbf{v}_x^d \in \mathbf{I}_1$) that $\mathbf{v}_y^d \in \mathbf{I}_2$ was detected within a radius r of \mathbf{v}_y . For each pair of images, we compute P_C twice by switching \mathbf{I}_1 and \mathbf{I}_2 and averaging the results.

The proposed Hybrid detector provides a fixed detection grid, while detectors such as SIFT detect a varying number of feature points per image, as they utilize a cornerness score that is compared with a detection threshold. For instance, the VIS-NIR dataset consists of 680×1024 images, resulting in 78×121 detection maps, while the SIFT typically detects ~ 500 feature points at most per image. Denote $N_s(I)$ as the number of SIFT-based feature points detected in an image I . A denser detection grid might result

Network/descriptor	VEDAI	CUHK	VIS-NIR
Engineered Features			
SIFT[25]	42.74	5.87	32.53
Inv SIFT [26]	11.33	7.34	27.71
LSS[33]	39.9	43.11	42.25
DASC[20]	8.9	43.05	38.1
LGHD[5]	1.31	0.65	10.76
Learning-based			
2Channel[3]	0	0.39	11.32
Q-Net 2P-4N[4]	0.78	0.9	22.5
Siamese[12]	0.84	3.38	13.17
Pseudo Siamese[12]	1.37	3.7	15.6
TS-Net[12]	0.45	2.77	11.86
Hybrid CNN			
Hybrid-CE	0.03	0.23	5.78
Hybrid-CE-HM	0	0.05	3.66
Hybrid-L_2	0	0.29	9.9
Hybrid-L_2-HM	0	0.1	3.41

Table 4. Patch matching results evaluated using the VIS-NIR, VEDAI and CUHK datasets, where the patches were extracted using a uniform lattice layout as in En et al. [12]. The accuracy is given in terms of the FPR95 score. The schemes names in bold are variations of the proposed scheme, while HM and CE relate to using hard mining and a Cross-Entropy loss, respectively.

in a higher detection probability P_C . Thus, to allow a fair comparison, we utilize $N_s(I)$ feature points per image for *all* detectors. The points in the Hybrid-based detection map are sorted by their activation, and the leading $N_s(I)$ feature points are retained.

We compared the proposed detector to the SIFT [25], SURF [7], KAZE [6], BRISK [22], and ORB [32] detectors. We also applied the proposed backtracking-based detection approach using the asymmetric subnetwork of the Hybrid CNN. All of these schemes were applied to the VEDAI [31], CUHK [39], and VIS-NIR [3] datasets. The detection results are shown in Fig. 3 where we report the cumulative detection probabilities as in Eq. 6. It follows that the proposed Hybrid detector significantly outperformed the hand-crafted detectors. Both the asymmetric and symmetric Hybrid subnetworks perform similarly, where the symmetric subnetwork performed best for the CUHK dataset consisting of corresponding face images and sketches. It is the dataset where the appearance differences are the most significant. For the VEDAI and VIS-NIR datasets, the appearance differences are less significant, and the asymmetric and symmetric subnetworks performed similarly.

5.4. Ablation study

We conducted an ablation study by comparing the results of the proposed Hybrid CNN to CNN formulations where one of the algorithmic components is omitted or changed. This allows evaluating the contribution of each of the pro-

Network	VIS-NIR
Hybrid-CE-SL-HM	4.72
Symmetric-CE-HM	4.2
Asymmetric-CE-HM	5.8
Hybrid-CE-ML	5.78
Hybrid-CE-ML-HM	3.66
Hybrid- L_2 -SL-HM	4.52
Symmetric- L_2 -HM	9.56
Asymmetric- L_2 -HM	4.32
Hybrid- L_2 -ML	9.90
Hybrid- L_2 -ML-HM	3.41

Table 5. Ablation results evaluated using the VIS-NIR dataset, where the patches were extracted using a uniform lattice layout as in En et al. [12]. The accuracy is given in terms of the FPR95 score and CE relates to applying the Cross-Entropy loss, while SL relates to training the CNN using a single loss L_H , as in Fig. 2. In contrast, ML relates to training using multiple losses and HM to using hard mining.

posed components. Thus, we first compare to using only the Siamese and Asymmetric CNNs and also compare the results of training the Hybrid CNN using the multi-task approach as in Section 3, to using a single loss (L_H in Fig. 2) while omitting the auxiliary losses L_S and L_A . We also show the added value of applying hard negative mining. The different CNNs were applied to the VIS-NIR [3] dataset using the experimental setup by En et al. [12], the same as in Section 5.2. The results are reported in Table 5, where the Hybrid CNN outperforms the Siamese and asymmetric CNNs, and that the multi-task learning allows to further improve both the training and validation losses for both L_2 and Cross-Entropy-based formulations. The hard mining (HM) also improves the patch classification accuracy significantly.

6. Conclusions

In this work, we presented a Deep-Learning approach for the detection and matching of feature points in multimodal images, that utilizes a novel Hybrid CNN formulation consisting of two CNN sub-networks. The first is a Siamese (weigh-sharing) CNN, while the second is an asymmetric (non-weigh-sharing) CNN. A novel feature points detection approach is derived by backtracking through the Siamese subnetwork, following the dominant activations. We show that the matching accuracy is improved by applying multi-task learning to the Siamese and asymmetric subnetworks, alongside the principal output loss. The proposed scheme is experimentally shown to outperform state-of-the-art approaches when applied to multiple multimodal image datasets. It significantly reduces the matchings errors by two to threefold and outperforms state-of-the-art detectors such as SIFT and SURF in terms of repeatability.

References

- [1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 1
- [2] Cristhian Aguilera, Fernando Barrera, Felipe Lumbrales, Angel D Sappa, and Ricardo Toledo. Multispectral image feature points. *Sensors*, 12(9):12661–12672, 2012. 1, 3
- [3] Cristhian A. Aguilera, Francisco J. Aguilera, Angel D. Sappa, Cristhian Aguilera, and Ricardo Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 9. IEEE, Jun 2016. 1, 2, 3, 5, 6, 7, 8
- [4] Cristhian A Aguilera, Angel D Sappa, Cristhian Aguilera, and Ricardo Toledo. Cross-spectral local descriptors via quadruplet network. *Sensors*, 17(4):873, 2017. 2, 3, 5, 6, 7, 8
- [5] C. A. Aguilera, A. D. Sappa, and R. Toledo. LGHD: A feature descriptor for matching across non-linear intensity variations. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 178–181, Sep. 2015. 1, 3, 6, 7, 8
- [6] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J. Davison. Kaze features. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 214–227, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 8
- [7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 5, 6, 8
- [8] Andrei Z. Broder. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, COM '00, pages 1–10, Berlin, Heidelberg, 2000. Springer-Verlag. 2, 4
- [9] Matthew Brown and Sabine Süsstrunk. Multi-spectral SIFT for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 177–184. IEEE, 2011. 6
- [10] Jian Chen and Jie Tian. Real-time multi-modal rigid registration based on a novel symmetric-SIFT descriptor. *Progress in Natural Science*, 19(5):643–651, 2009. 1, 2
- [11] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9. IEEE, Jun 2019. 3, 5
- [12] S. En, A. Lechervy, and F. Jurie. TS-NET: Combining modality specific and common features for multimodal patch matching. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3024–3028, Oct 2018. 2, 3, 5, 6, 7, 8
- [13] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. *ArXiv e-prints*, July 2017. 5
- [14] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases, VLDB '99*, pages 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. 2, 4
- [15] Mahmudul Hasan, Mark R Pickering, and Xiuping Jia. Modified SIFT for multi-modal remote sensing image registration. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pages 2348–2351. IEEE, 2012. 1, 3, 5
- [16] Md Tanvir Hossain, Guohua Lv, Shyh Wei Teng, Guojun Lu, and Martin Lackmann. Improved symmetric-SIFT for multi-modal image registration. In *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*, pages 197–202. IEEE, 2011. 1, 3
- [17] Michal Irani and P Anandan. Robust multi-sensor image alignment. In *Computer Vision, 1998. Sixth International Conference on*, pages 959–966. IEEE, 1998. 1, 2, 5
- [18] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *ArXiv e-prints*, Dec. 2014. 5
- [19] Yosi Keller and Amir Averbuch. Multisensor image registration via implicit similarity. *IEEE transactions on pattern analysis and machine intelligence*, 28(5):794–801, 2006. 1, 2, 5
- [20] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2103–2112, 2015. 3, 6, 7, 8
- [21] Youngwook P Kwon, Hyojin Kim, Goran Konjevod, and Sara McMains. Dude (duality descriptor): A robust descriptor for disparate images using line segment duality. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 310–314. IEEE, 2016. 1, 3
- [22] S. Leutenegger, M. Chli, and R. Y. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *2011 International Conference on Computer Vision*, pages 2548–2555, Nov 2011. 8
- [23] Hui Li, BS Manjunath, and Sanjit K Mitra. A contour-based approach to multisensor image registration. *IEEE transactions on image processing*, 4(3):320–334, 1995. 1, 5
- [24] Tony Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318, Dec 1993. 2
- [25] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1, 2, 4, 5, 6, 7, 8
- [26] Rui Ma, Jian Chen, and Zhong Su. MI-SIFT: Mirror and inversion invariant generalization for SIFT descriptor. In *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '10*, pages 228–235, New York, NY, USA, 2010. ACM. 2, 6, 7, 8
- [27] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, Oct 2005. 4, 5, 7

- [28] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenović, and Jiří Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pages 4829–4840, USA, 2017. Curran Associates Inc. 3, 6
- [29] Dou Quan, Shuai Fang, Xuefeng Liang, Shuang Wang, and Licheng Jiao. Cross-spectral image patch matching by learning features of the spatially connected patches in a shared space. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 115–130, Cham, 2019. Springer International Publishing. 3, 6, 7
- [30] Dou Quan, Xuefeng Liang, Shuang Wang, Shaowei Wei, Yanfeng Li, Ning Huyan, and Licheng Jiao. AFD-Net: Aggregated feature difference learning for cross-spectral image patch matching. In *Proceedings of the 2019 International Conference on Computer Vision, ICCV '19*, Washington, DC, USA, 2019. IEEE Computer Society. 2, 3
- [31] Sébastien Razakarivony and Frédéric Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016. 6, 7, 8
- [32] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the 2011 International Conference on Computer Vision, ICCV '11*, pages 2564–2571, Washington, DC, USA, 2011. IEEE Computer Society. 6, 8
- [33] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 3, 6, 7, 8
- [34] O. Simeoni, Y. Avrithis, and O. Chum. Local features and visual words emerge in activations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9. IEEE, Jun 2019. 3, 5
- [35] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013. 1
- [36] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2010. 1
- [37] Daniel Vaquero, Matthew Turk, Kari Pulli, Marius Tico, and Natasha Gelfand. A survey of image retargeting techniques. In *Proc. SPIE*, volume 7798, page 779814, 2010. 1
- [38] Paul Viola and William M. Wells, III. Alignment by maximization of mutual information. *Int. J. Comput. Vision*, 24(2):137–154, Sept. 1997. 1, 2
- [39] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009. 6, 7, 8
- [40] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2015. 2, 3
- [41] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003. 1