Semantic Attribute Matching Networks

Seungryong Kim[1,2], Dongbo Min[3], Somi Jeong[1], Sunok Kim[1,2], Sangryul Jeon[1], Kwanghoon Sohn[1,*]

[1]Yonsei University, [2]École Polytechnique Fédérale de Lausanne (EPFL), [3]Ewha Womans University

seungryong.kim@epfl.ch

{somijeong,kso428,cheonjsr,khsohn}@yonsei.ac.kr, dbmin@ewha.ac.kr

Abstract

We present semantic attribute matching networks (SAM-Net) for jointly establishing correspondences and transferring attributes across semantically similar images, which intelligently weaves the advantages of the two tasks while overcoming their limitations. SAM-Net accomplishes this through an iterative process of establishing reliable correspondences by reducing the attribute discrepancy between the images and synthesizing attribute transferred images using the learned correspondences. To learn the networks using weak supervisions in the form of image pairs, we present a semantic attribute matching loss based on the matching similarity between an attribute transferred source feature and a warped target feature. With SAM-Net, the state-of-the-art performance is attained on several benchmarks for semantic matching and attribute transfer.

1. Introduction

Establishing correspondences and transferring attributes across *semantically* similar images can facilitate a variety of computer vision applications [35, 34, 25]. In these tasks, the images resemble each other in contents but differ in visual attributes, such as color, texture, and style, e.g., the images with different faces as exemplified in Fig. 1. Numerous techniques have been proposed for the semantic correspondence [15, 24, 42, 19, 43, 23] and attribute transfer [11, 6, 28, 21, 38, 16, 20, 16, 34, 12], but these two tasks have been studied independently although they can be mutually complementary.

To establish reliable semantic correspondences, state-of-the-art methods have leveraged deep convolutional neural networks (CNNs) in extracting descriptors [7, 53, 24] and regularizing correspondence fields [15, 42, 19, 43, 23].
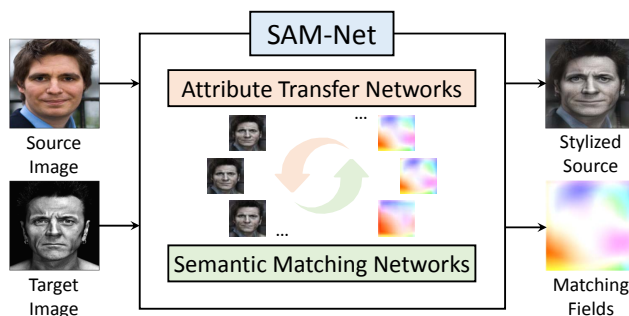
Figure 1. Illustration of SAM-Net: for semantically similar images having both photometric and geometric variations, SAM-Net recurrently estimates semantic correspondences and synthesizes attribute transferred images in a joint and boosting manner.

Compared to conventional handcrafted methods [35, 22, 5, 54, 48], they have achieved a highly reliable performance. To overcome the problem of limited ground-truth supervisions, some methods [42, 19, 43, 23] have tried to learn deep networks using only weak supervision in the form of image pairs based on the intuition that the matching cost between the source and target features over a set of transformations should be minimized at the correct transformation. These methods presume that the attribute variations between source and target images are negligible in the deep feature space. However, in practice the deep features often show limited performance in handling different attributes that exist in the source and target images, often degrading the matching accuracy dramatically.

To transfer the attributes between source and target images, following the seminal work of Gatys et al. [10], numerous methods have been proposed to separate and recombine the contents and attributes using deep CNNs [11, 6, 28, 21, 38, 16, 20, 16, 34, 12]. Unlike the parametric methods [11, 21, 38, 16] that match the global statistics of deep features while ignoring the spatial layout of contents, the non-parametric methods [6, 28, 34, 12] directly find neural patches in the target image similar to the source patch and synthesize them to reconstruct the stylized image. These non-parametric methods generally estimate nearest

neighbor patches between source and target images with weak implicit regularization methods [6, 28, 34, 12] using a simple local aggregation followed by a winner-takes-all (WTA). However, photorealistic attribute transfer needs highly regularized and semantically meaningful correspondences, and thus existing methods [6, 28, 12] frequently fail when the images have background clutters and different attributes while representing similar global feature statistics. A method called deep image analogy [34] has tried to estimate more semantically meaningful dense corrrespondences for photorealistic attribute transfer, but it still has limited localization ability with PatchMatch [3].

In this paper, we present semantic attribute matching networks (SAM-Net) for overcoming the aforementioned limitations of current semantic matching and attribute transfer techniques. The key idea is to weave the advantages of semantic matching and attribute transfer networks in a boosting manner. Our networks accomplish this through an iterative process of establishing more reliable semantic correspondences by reducing the attribute discrepancy between semantically similar images and synthesizing an attribute transferred image with the learned semantic correspondences. Moreover, our networks are learned from weak supervision in the form of image pairs using the proposed semantic attribute matching loss. Experimental results show that SAM-Net outperforms the latest methods for semantic matching and attribute transfer on several benchmarks, including TSS dataset [48], PF-PASCAL dataset [14], and CUB-200-2011 dataset [51].

2. Related Work

Semantic correspondence. Most conventional methods for semantic correspondence that use handcrafted features and regularization methods [35, 22, 5, 54, 48] have provided limited performance due to a low discriminative power. Recent approaches have used deep CNNs for extracting their features [7, 53, 24, 39] and regularizing correspondence fields [15, 41, 42]. Rocco et al. [41, 42] proposed deep architecture for estimating a geometric matching model, but these methods estimate only globally-varying geometric fields. To deal with locally-varying geometric deformations, some methods such as UCN [7] and CAT-FCSS [25] were proposed based on STNs [18]. Recently, PARN [19], NC-Net [43], and RTNs [23] were proposed to estimate locally-varying transformation fields using a coarse-to-fine scheme [19], neighbourhood consensus [43], and an iteration technique [23]. These methods [19, 43, 23] presume that the attribute variations between source and target images are negligible in the deep feature space. However, in practice the deep features often show limited performance in handling different attributes. Aberman et al. [1] presented a method to deal with the attribute variations between the images using a variant of instance normalization [16]. How-

ever, the method does not have an explicit learnable module to reduce the attribute discrepancy, thus yielding limited performance.

Attribute transfer. There have been a lot of works on the transfer of visual attributes, e.g., color, texture, and style, from one image to another, and most approaches are tailored to their specific objectives [40, 47, 8, 2, 52, 9]. Since our method represents and synthesizes deep features to transfer the attribute between semantically similar images, the neural style transfer [11, 6, 21, 20] is highly related to ours. In general, these approaches can be classified into parametric and non-parametric methods.

In parametric methods, inspired by the seminal work of Gatys et al. [10], numerous methods have been presented, such as the work of Johnson et al. [21], AdaIN [16], and WCT [31]. Since these methods are globally formulated, they have shown limited performance for photorealistic stylization tasks [32, 38]. To alleviate these limitations, Luan et al. proposed a deep photo style transfer [38] that computes and uses the semantic labels. Li et al. proposed Photo-WCT [32] to eliminate the artifacts using additional smoothing step. However, these methods still have been formulated without considering semantically meaningful correspondence fields.

Among non-parametric methods, the seminal work of Li et al. [28] first searches local neural patches, which are similar to the patch of content image, in the target style image to preserve the local structure prior of content image, and then uses them to synthesize the stylized image. Chen et al. [6] sped up this process using the feed-forward networks to decode the synthesize features. Inspired by this, various approaches have been proposed to synthesize locally blended features efficiently [29, 49, 37, 30, 50]. However, the aforementioned methods are tailored to the artistic style transfer, and thus they focused on finding the patches to reconstruct more plausible images, rather than finding semantically meaningful dense correspondences. They generally estimate the nearest neighbor patches using weak implicit regularization methods such as WTA. Recently, Gu et al. [12] introduced a deep feature reshuffle technique to connect both parametric and non-parametric methods, but they search the nearest neighbor using an expectation-maximization (EM) that also produces limited localization accuracy.

More related to our work is a method called deep image analogy [34] that searches semantic correspondences using deep PatchMatch [3] in a coarse-to-fine manner. However, PatchMatch inherently has a limited regularization power as shown in [27, 36, 33]. In addition, the method still needs the greedy optimization for feature deconvolution that induces computational bottlenecks, and only considers the translational fields, thus having the limitation to handle more complicated deformations.
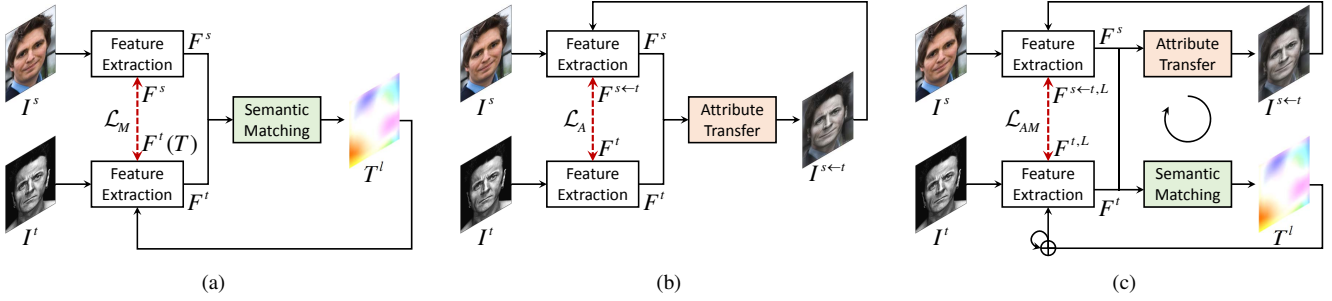
Figure 2. Intuition of SAM-Net: (a) methods for semantic matching [41, 42, 23, 19], (b) methods for attribute transfer [11, 21, 28], and (c) SAM-Net, which recurrently weaves the advantages of both existing semantic matching and attribute transfer techniques.

3. Problem Statement

Let us denote *semantically* similar source and target images as I^s and I^t, respectively. The objective of our method is to jointly establish a correspondence field $f_i = [u_i, v_i]^T$ between the two images that is defined for each pixel $i = [i_x, i_y]^T$ and synthesize an attribute transferred image $I^{s \leftarrow t}$ by transferring an attribute of target image I^t to a content of source image I^s.

CNN-based methods for semantic correspondence [41, 25, 42, 19, 43, 23] involve first extracting deep features [45, 25], denoted by F_i^s and F_i^t, from I_i^s and I_i^t within local receptive fields, and then estimating correspondence field f_i of the source image using deep regularization models [41, 42, 23], as shown in Fig. 2(a). To learn the networks using only image pairs, some methods [42, 23] formulate the loss function based on the intuition that the matching cost between the source feature F_i^s and the target feature $F_{i+f_i}^t$ over a set of transformations should be minimized. For instance, they formulate the matching loss defined as

$$\mathcal{L}_M = \sum\nolimits_i \|F_i^s - F_{i+f_i}^t\|_F^2, \tag{1}$$

where $\| \cdot \|_F^2$ denotes Frobenius norm. To deal with more complex deformations such as affine transformation [27, 23], instead of $F_{i+f_i}^t$, $F^t(T_i)$ or $F_{i+f_i}^t(A_i)$ can be used with a 2×3 matrix $T_i = [A_i, f_i]$. Although semantically similar images can share similar contents but have different attributes, these methods [41, 42, 19, 43, 23] simply assume that the attribute variations between source and target images are negligible in the deep feature space. It thus cannot guarantee measuring a fully accurate matching cost without an explicit module to reduce the attribute gaps.

To minimize the attribute discrepancy between source and target images, attribute or style transfer methods [11, 6, 21, 20] separate and recombine the content and attribute. Unlike the parametric methods [11, 38], the non-parametric methods [6, 28, 34, 12] directly find neural patches in the target image similar to the source patch and synthesize them to reconstruct the stylized feature $F^{s \leftarrow t}$ and image $I^{s \leftarrow t}$, as shown in Fig. 2(b). Formally, they formulate two loss func-

tions including the content loss defined as

$$\mathcal{L}_C = \sum\nolimits_i \|F_i^{s \leftarrow t} - F_i^s\|_F^2, \tag{2}$$

and the non-parametric attribute transfer loss defined as

$$\mathcal{L}_A = \sum\nolimits_i \sum\nolimits_{j \in N_i} \|F_j^{s \leftarrow t} - F_{j+f_i}^t\|_F^2, \tag{3}$$

where $i + f_i$ is the center point of the patch in I^t that is most similar to a patch centered at i in I^s. Generally, f_i is determined using the matching scores of normalized cross-correlation [6, 28] aggregated on N_i over all local patches followed by the labeling optimization such that

$$f_i = \underset{m}{\operatorname{argmax}} \sum\nolimits_{j \in N_i} (F_j^s \cdot F_{j+m}^t) / \|F_j^s\| \|F_{j+m}^t\|, \tag{4}$$

where the operator $\cdot$ denotes inner product.

However, the hand-designed discrete labeling techniques such as WTA [6, 28], PatchMatch [34], and EM [12] used to optimize (4) rely on weak implicit smoothness constraints, often producing poor matching results. In addition, they only consider the translational fields, i.e., f_i, thus limiting handling more complicated deformations caused by scale, rotation and skew that may exist among object instances.

4. Method

4.1. Overview

We present the networks to recurrently estimate semantic correspondences and synthesize the stylized images in a boosting manner, as shown in Fig. 2(c). In the networks, correspondences are robustly established by matching the stylized source and target images, in contrast to existing methods [42, 23] that directly match source and target images that have the attribute discrepancy. At the same time, blended neural patches using the correspondences are used to reconstruct the attribute transferred image in a semantic-aware and geometrically aligned manner.

Our networks are split into three parts as shown in Fig. 3: *feature extraction networks* to extract source and target features F^s and F^t, *semantic matching networks* to establish
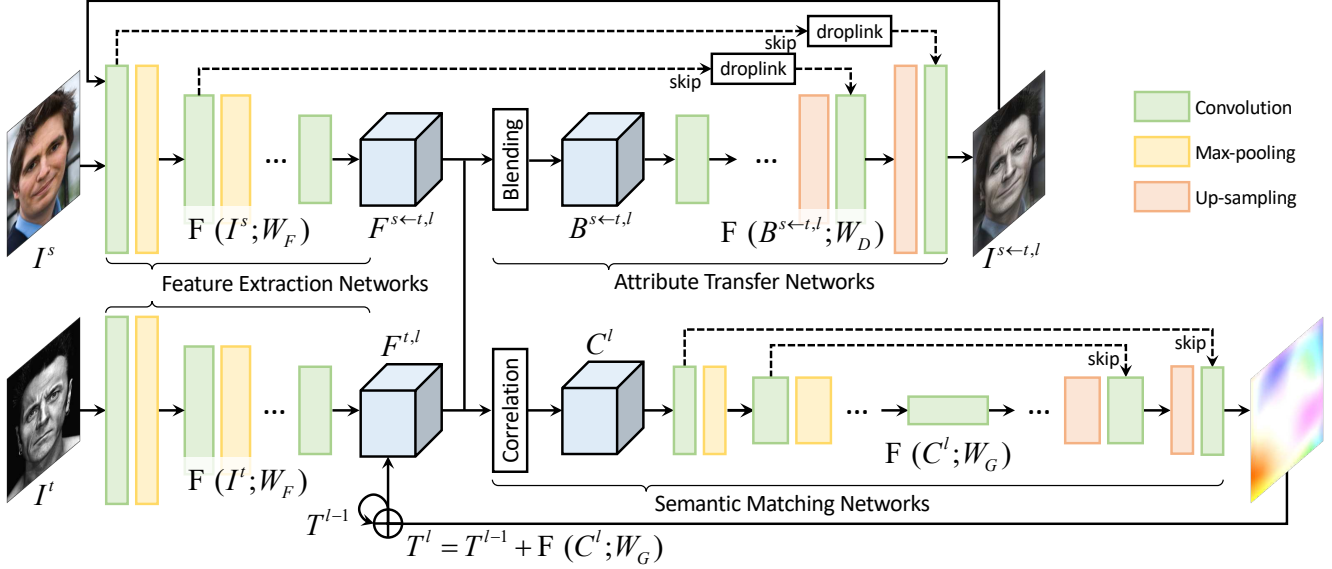
Figure 3. Network configuration of SAM-Net, consisting of feature extraction networks, semantic matching networks, and attribute transfer networks in a recurrent structure. Initially, $F^{s \leftarrow t,0} = F^s$ and $F^{t,0} = [I_{2 \times 2}, 0_{2 \times 1}]$. They output T_i^l and $I^{s \leftarrow t,l}$ at each l-th iteration.

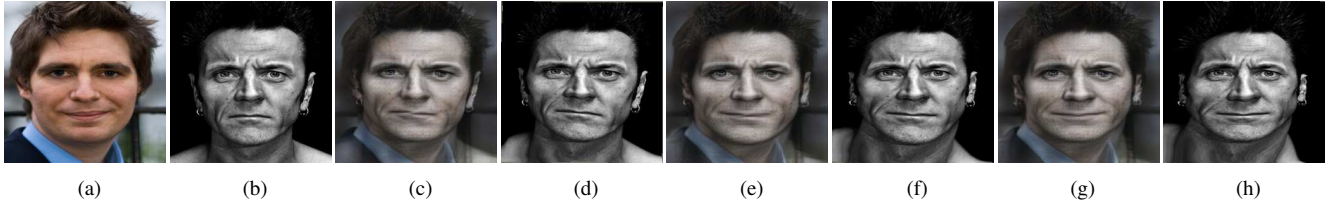| (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) |

Figure 4. Convergence of SAM-Net: (a) source image, (b) target image, iterative evolution of attribute transferred images (c), (e), and (g) and warped images using dense corresondences (d), (f), and (h) after iteration 1, 2, and 3. In the recurrent formulation of SAM-Net, the predicted transformation fields and attribute transferred images become progressively more accurate through iterative estimation.

correspondence fields T, and *attribute transfer networks* to synthesize the attribute transferred image $I^{s \leftarrow t}$. Since our networks are formulated in a recurrent manner, they output T^l and $I^{s \leftarrow t,l}$ at each l-th iteration, as exemplified in Fig. 4.

4.2. Network Architecture

Feature extraction networks. Our model accomplishes the semantic matching and attribute transfer using deep features [45, 25]. To extract the features for source F^s and target F^t, the source and target images (I^s and I^t) are first passed through shared feature extraction networks with parameters W_F such that $F_i = \mathcal{F}(I_i; W_F)$, respectively. In the recurrent formulation, an attribute transferred feature $F^{s \leftarrow t,l}$ from target to source images and a warped target feature $F^{t,l}$, i.e., F^t warped using the transformation fields T_i^l, are reconstructed at each l-th iteration.

Semantic matching networks. Our semantic matching networks consist of the matching cost computation and inference modules motivated by conventional RANSAC-like methods [17]. We first compute the correlation volume with respect to translational motion only [41, 42, 43, 23] and then pass it to subsequent convolutional layers to determine

dense affine transformation fields T_i.

Unlike existing methods [41, 42, 23], our method computes the matching similarity between not only source and target features but also synthesized source and target features to minimize errors from the attribute discrepancy between source and target features such that:

$$C_i^l(p) = (1 - \lambda^l)(F_i^s \cdot F_p^{t,l})/\|F_i^s\| \|F_p^{t,l}\| + \lambda^l (F_i^{s \leftarrow t,l} \cdot F_p^{t,l})/\|F_i^{s \leftarrow t,l}\| \|F_p^{t,l}\|, \quad (5)$$

where $p \in P_i$ for local search window P_i centered at i. λ^l controls the trade-off between content and attribute when computing the similarity, which is similar to [34]. Note that when $\lambda^l = 0$, we only consider the source feature F^s without considering the stylized feature $F^{s \leftarrow t}$. These similarities undergo L_2 normalization to reduce errors [42].

Based on this, the matching inference networks with parameters W_G iteratively estimate the residual between the previous and current transformation fields [23] as

$$T_i^l - T_i^{l-1} = \mathcal{F}(C_i^l; W_G). \quad (6)$$

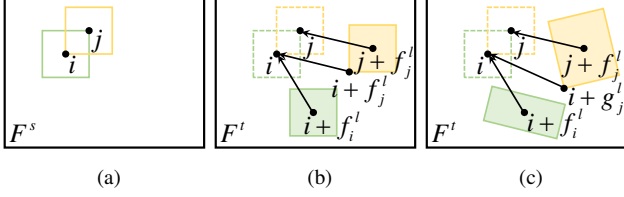The current transformation fields are then estimated in a re-

Figure 5. Visualization of neural patch blending: for source feature F^s in (a), unlike existing methods [34, 28, 12] that blend features of source F^s and target F^t using only traslationional fields f_i as in (b), our method blends the features with the learned affine transformation fields $T_i^l = [A_i^l, f_i^l]$ as in (c).

Figure 6. Effects on the confidence in neural patch blending: (a) blending results of I^s and I^t, (b) blending results of F^s and F^t followed by the decoder, (c) confidence, and (d) blending results of F^s and F^t with the confidence followed by the decoder.

current manner [23] as follows:

$$T_i^l = [I_{2\times2}, 0_{2\times1}] + \sum_{n \in \phi(l)} \mathcal{F}(C_i^n; W_G), \qquad (7)$$

where $\phi(l) = \{1, .., l-1\}$. Unlike [41, 42] that estimate a global affine or thin-plate spline transformation field, our networks are formulated as the encoder-decoder networks as in [44] to estimate locally-varying transformation fields.

Attribute transfer networks. To transfer the attribute of target feature F^t into the content of source feature F^s at l-th iteration, our attribute transfer networks first blend the source and target features as $B^{s \leftarrow t,l}$ using estimated transformation field T_i^l and then reconstruct the stylized source image $I^{s \leftarrow t,l}$ using the decoder networks with parameters W_D such that $I^{s \leftarrow t,l} = \mathcal{F}(B^{s \leftarrow t,l}; W_D)$.

Specifically, our neural patch blending between F^s and F^t with the current transformation field $T_i^l = [A_i^l, f_i^l]$ is formulated as shown in Fig. 5 such that

$$B_i^{s \leftarrow t,l} = (1 - \lambda^l)F_i^s + \lambda^l \sum_{j \in N_i} \alpha_j^l F_{i+g_j^l}^t / \sum_{j \in N_i} \alpha_j^l, \qquad (8)$$

where $g_j^l = (A_j^l - I_{2\times2})(i-j) + f_j^l$. α_i^l is a confidence of each pixel i that has T_i^l computed similar to [26] such that

$$\alpha_i^l = \exp(C_i^l(i)) / \sum_{p \in P_i} \exp(C_i^l(p)). \qquad (9)$$

Our neural patch blending module differs from the existing methods [34, 28, 12] in the use of learned transformation fields and consideration of more complex deformations such as affine transformations. In addition, unlike exisiting style transfer methods [28, 12], our networks employ the confidence to transfer the attribute of matchable points only tailored to our objective, as exemplified in Fig. 6.

In addition, our decoder networks are formulated as a symmetric structure to feature extraction networks. Since the single-level decoder networks as in [16] cannot capture both complicated structures at high-level features and low-level information at low-level features, the multi-level decoder networks have been proposed as in [31, 32], but they are not very economic [12]. Instead, we use the skip connection from the source features F^s to capture both low-
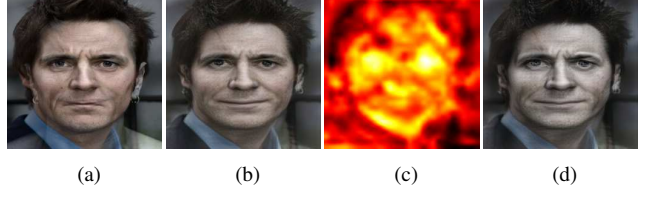
and high-level attribute characteristics [31, 32, 12]. However, using the skip connection through simple concatenation [44] makes the decoder networks reconstruct an image using only low-level features. To alleviate this, inspired by a dropout layer [46], we present a droplink layer such that the skipped features and upsampled features are stochastically linked to avoid the overfitting to certain level features:

$$F_h^{s \leftarrow t,l} = (1 - b_h)\mathcal{F}(B^{s \leftarrow t,l}; W_{D,h}) + b_h F_h^s, \qquad (10)$$

where $F_h^{s \leftarrow t,l}$ and F_h^s are the intermediate and skipped features at h-th level for $h \in \{1, ..., H\}$. $W_{D,h}$ is the parameters until h-th level. b_h is a binary random variable. Note that if $b_h = 0$, this becomes the no-skip connected layer.

4.3. Loss Functions

Semantic attribute matching loss. Our networks are learned using weak supervision in the form of image pairs. Concretely, we present a semantic attribute matching loss in a manner that the transformation field T and the stylized image $I^{s \leftarrow t}$ can be simultaneously learned and inferred to minimize a single loss function. After the convergence of iterations at L-th iteration, an attribute transferred feature $F^{s \leftarrow t,L}$ and a warped target feature $F^{t,L}$ are used to define the loss function. This intuition can be realized by minimizing the following objective:

$$D(F^{s \leftarrow t,L}, F^{t,L}) = \sum_i \sum_{j \in N_i} \|F_j^{s \leftarrow t,L} - F_j^{t,L}\|_F^2. \quad (11)$$

In comparison to existing the matching loss $\mathcal{L}_M$ and the attribute transfer loss $\mathcal{L}_A$, this objective enables us to solve the photometric and geometric variations across semantically similar images simultaneously.

Although using only this objective provides satisfactory performance, we extend this objective to consider both positive and negative samples to enhance network training and precise localization ability based on the intuition that the matching score should be minimized at the correct transformation while keeping the scores of other neighbor transformation candidates high. Finally, we formulate our semantic attribute matching loss as a cross-entropy loss as

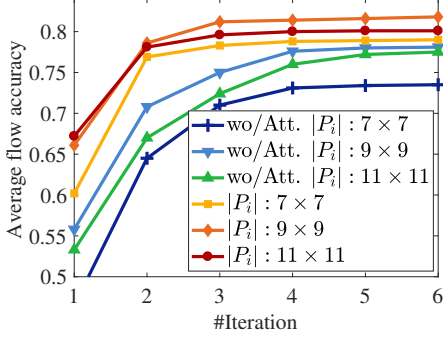$$\mathcal{L}_{AM} = \sum_i \max(-\log(K_i), \tau), \qquad (12)$$

Figure 7. Convergence analysis of SAM-Net for various numbers of iterations and search window sizes on the TSS benchmark [48].

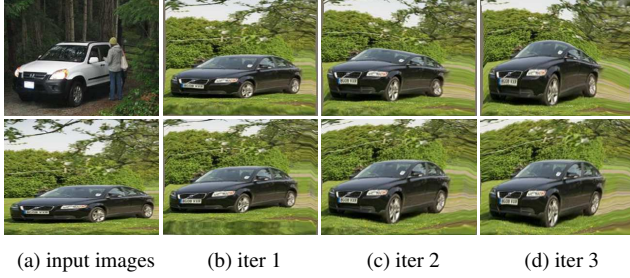| (a) input images | (b) iter 1 | (c) iter 2 | (d) iter 3 |

Figure 8. Ablation study of SAM-Net without (top) and with (bottom) attribute transfer networks as evolving iterations.

where K_i is the softmax probability defined as

$$K_i = \frac{\exp(-D(F_i^{s \leftarrow t, L}, F_i^{t, L}))}{\sum_{q \in Q_i} \exp(-D(F_i^{s \leftarrow t, L}, F_q^{t, L}))}. \quad (13)$$

It makes the center point i within the neighbor Q_i become a positive sample and the other points become negative samples. In addition, the truncated max operator $\max(\cdot, \tau)$ is used to focus on the sailent parts such as objects during training with the parameter τ.

Other losses. We utilize two additional losses, namely the content loss $\mathcal{L}_C$ as in (2) to preserve the structure of source image and the L_2 regularization loss [21, 28] to encourage spatial smoothness in the stylized image.

5. Experiments

5.1. Training and Implementation Details

To learn our SAM-Net, large-scale semantically similar image pairs are needed, but such public datasets are limited quantitatively. To overcome this, we adopt a two-step training technique, similar to [42]. In the first step, we train our networks using a synthetic training dataset provided in [41], where synthetic transformations are randomly applied to a single image to generate the image pairs, and thus the images do not have appearance variations. This enables the attribute transfer networks to be learned in an auto-encoder manner [31, 16, 32], but the matching networks still have

Methods	FG3D	JODS	PASC.	Avg.
Taniai et al. [48]	0.830	0.595	0.483	0.636
PF [13]	0.786	0.653	0.531	0.657
DCTM [27]	0.891	0.721	0.610	0.740
SCNet [15]	0.776	0.608	0.474	0.619
GMat. [41]	0.835	0.656	0.527	0.673
GMat. w/Inl. [42]	0.892	0.758	0.562	0.737
DIA [34]	0.762	0.685	0.513	0.653
RTNs [23]	0.901	0.782	0.633	0.772
SAM-Net w/(11)	0.891	0.789	0.638	0.773
SAM-Net wo/Att.	0.912	0.790	0.641	0.781
SAM-Net	**0.961**	**0.822**	**0.672**	**0.818**

Table 1. Matching accuracy compared to the state-of-the-art correspondence techniques on the TSS benchmark [48].

Methods	PCK		
	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$
PF [13]	0.314	0.625	0.795
DCTM [27]	0.342	0.696	0.802
SCNet [15]	0.362	0.722	0.820
GMat. [41]	0.410	0.695	0.804
GMat. w/Inl. [42]	0.490	0.748	0.840
DIA [34]	0.471	0.724	0.811
RTNs [23]	0.552	0.759	0.852
NC-Net [43]	-	0.789	-
SAM-Net	**0.601**	**0.802**	**0.869**

Table 2. Matching accuracy compared to the state-of-the-art correspondence techniques on the PF-PASCAL benchmark [14].

limited ability to deal with the attribute variations. To overcome this, in the second step, we finetune this pretrained network on public datasets for semantically similar image pairs from the training set of PF-PASCAL [14] following the split used in [14].

For feature extraction, we used the ImageNet-pretrained VGG-19 networks [45], where the activations are extracted from 'relu4-1' layer (i.e., $H = 4$). We gradually increase λ^l until 1 such that $\lambda^l = 1 - \exp(-l)$. During training, we set the maximum number of iteration L to 5 to avoid the gradient vanishing and exploding problem. During testing, the iteration count is increased to 10. Following [23], the window sizes of N_i, P_i, and Q_i are set to 3×3, 9×9, and 9×9, respectively. The probability of b_h is defined as 0.9 and in testing b_h is set to 0.5.

5.2. Experimental Settings

In the following, we comprehensively evaluated SAM-Net through comparisons to state-of-the-art methods for semantic matching, including Taniai et al. [48], PF [13], SCNet [15], DCTM [24], DIA [34], GMat. [41], GMat. w/Inl. [42], NC-Net [43], RTNs [23], and for attribute transfer, including Gatys et al. [10], CNN-MRF [28], Photo-WCT [32], Gu et al. [12], and DIA [34]. Performance was

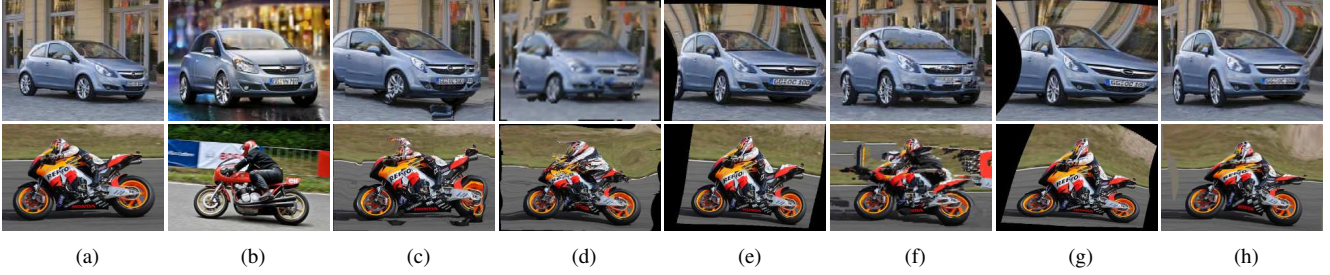| (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) |

Figure 9. Qualitative results on the TSS benchmark [48]: (a) source and (b) target images, warped source images using correspondences of (c) PF [13], (d) DCTM [27], (e) GMat [41], (f) DIA [34], (g) GMat. w/Inl. [42], and (h) SAM-Net.

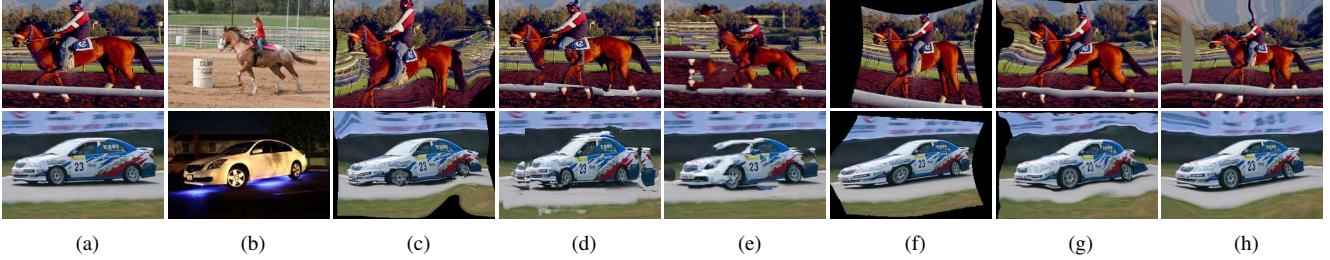| (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) |

Figure 10. Qualitative results on the PF-PASCAL benchmark [13]: (a) source and (b) target images, warped source images using correspondences of (c) DCTM [27], (d) SCNet [15], (e) DIA [34] (f) GMat. w/Inl. [42], (g) RTNs [23], and (h) SAM-Net.

measured on TSS dataset [48], PF-PASCAL dataset [14], and CUB-200-2011 dataset [51].

In Sec. 5.3, we first analyzed the effects of the components within SAM-Net, and then evaluated matching results with various benchmarks and quantitative measures in Sec. 5.4. We finally evaluated photorealistic attribute transfer results with various applications in Sec. 5.5.

5.3. Ablation Study

To validate the components within SAM-Net, we evaluated the matching accuracy for different numbers of iterations, with various sizes of P_i, and with and without attribute transfer module. For quantitative assessment, we examined the accuracy on the TSS benchmark [48]. As shown in Fig. 7, Fig. 8, and Table 1, SAM-Net converges in $2-3$ iterations. In addition, the results of 'SAM-Net wo/Att.', i.e., SAM-Net without attribute transfer, show the effectiveness of attribute transfer module in the recurrent formulation. The results of 'SAM-Net wo/(11).', i.e., SAM-Net with the loss of (11), show the importance to consider the negative samples when training. By enlarging the size of P_i, the accuracy improves until 9×9, but larger window sizes reduce matching accuracy due to greater matching ambiguity. Note that $Q_i = P_i$ following to [23].

5.4. Semantic Matching Results

TSS benchmark. We evaluated SAM-Net on the TSS benchmark [48], consisting of 400 image pairs. As in [24, 27], flow accuracy was measured in Table 1. Fig. 9 shows qualitative results. Unlike existing methods [7, 48, 13, 15, 24, 41, 42, 23] that do not consider the attribute varia-

tions between semantically similar images, our SAM-Net has shown highly improved preformance qualitatively and quantitatively. DIA [34] has shown limited matching accuracy compared to other deep methods [42, 23], due to their limited regularization powers. Unlike this, the results of our SAM-Net shows that our method is more successfully transferring the attribute between source and target images to improve the semantic matching accuracy.

PF-PASCAL benchmark. We also evaluated SAM-Net on the PF-PASCAL benchmark [14], which contains 1,351 image pairs over 20 object categories with PASCAL keypoint annotations [4]. For the evaluation metric, we used the PCK between flow-warped keypoints and the ground truth as done in the experiments of [15]. Table 2 summarizes the PCK values, and Fig. 10 shows qualitative results. Similar to the experiments on the TSS benchmark [48], CNN-based methods [15, 41, 42, 42, 23] including our SAM-Net yield better performance, with SAM-Net providing the highest matching accuracy.

5.5. Applications

Photorealistic attribute transfer. We evaluated SAM-Net for photorealistic attribute transfer on the TSS [48] and PF-PASCAL benchmarks [14]. For evaluatation, we sampled the image pairs from these datasets and transferred the attribute of target image to the source image as shown in Fig. 11. Note that SAM-Net is designed to work on images contain that semantically similar contents and not effective for generic artistic style transfer applications as in [10, 21, 16]. As expected, existing methods tailored to

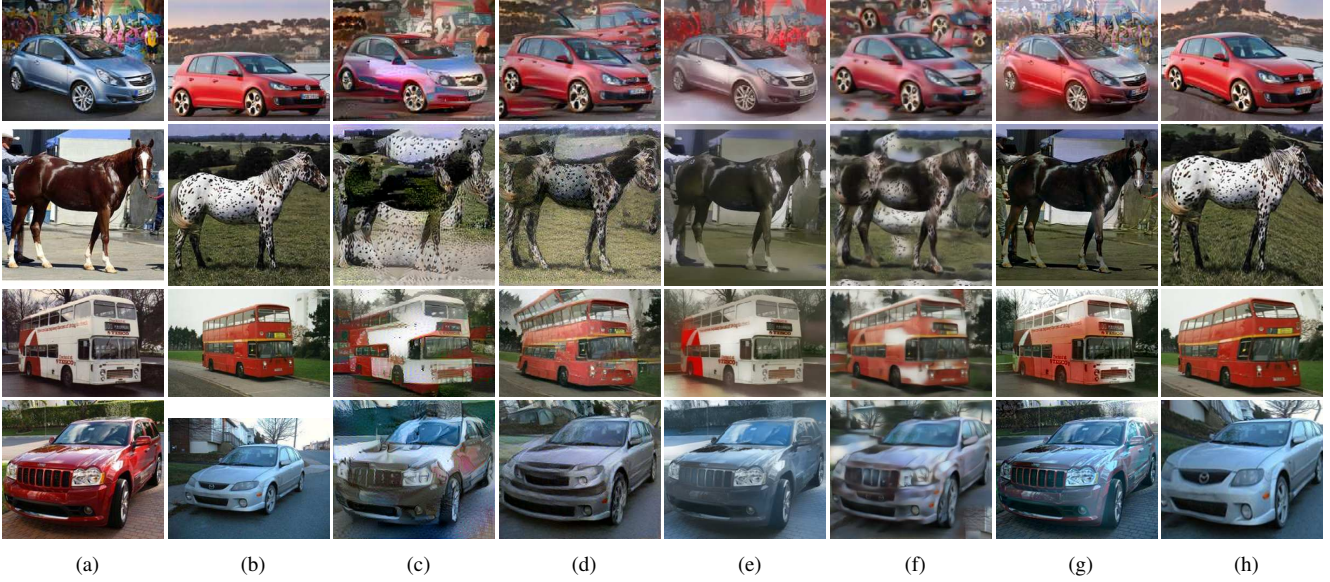(a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)

Figure 11. Qualitative results of the photorealistic attribute transfer on the TSS [48] PF-PASCAL [14] benchmarks: (a) source and (b) target images, results of (c) Gatys et al. [10], (d) CNN-MRF [28], (e) Photo-WCT [32], (f) Gu et al. [12], (g) DIA [34], and (h) SAM-Net.
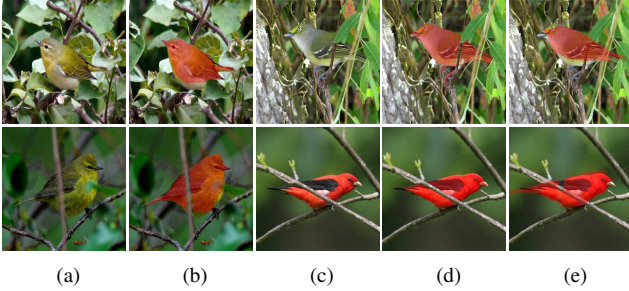

(a)   (b)   (c)   (d)   (e)

Figure 12. Qualitative results of the mask transfer on the CUB-200-2011 benchmark [51]: source (a) images and (b) masks and target (c) images and (d) masks, and (e) warped source masks to the target images using correspondences from SAM-Net.
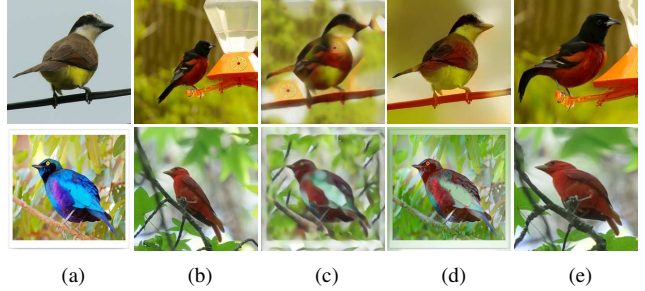

(a)   (b)   (c)   (d)   (e)

Figure 13. Qualitative results of the object transfiguration on the CUB-200-2011 benchmark [51]: (a) source and (b) target images, results of (c) Gu et al. [12], (d) DIA [34], and (e) SAM-Net.

artistic stylization such as a method of Gatys et al. [10] and CNN-MRF [28] produce limited quality images. Moreover, recent photorealistic stylization methods such as Photo-WCT [32] and Gu et al. [12] have limited performance for the images that have background clutters. DIA [34] provided degraded results due to its weak regularization technique. Unlike these methods, our SAM-Net has shown highly accurate and plausible results thanks to their learned transformation fields to synthesize the images. Note that some methods such as Photo-WCT [32] and DIA [34] have used to refine their results using additional smoothing modules, but SAM-Net does not use any post-processing.

Foreground mask transfer. We evaluated SAM-Net for mask transfer on the CUB-200-2011 dataset [51], which contains images of 200 bird categories, with annotated foreground masks. For semantically similar images that have very challenging photometric and geometric varia-

tions, our SAM-Net successfully transfers the semantic labels, as shown in Fig. 12.

Object transfiguration. We finally applied our method to object transfiguration, e.g., translating a source bird into a target breed. We used object classes from the CUB-200-2011 dataset [51]. In this application, our SAM-Net has shown very plausible results as exemplified in Fig. 13.

6. Conclusion

We presented SAM-Net that recurrently estimates dense correspondences and transfers the attributes across semantically similar images in a joint and boosting manner. The key idea of this approach is to formulate the semantic matching and attribute transfer networks to complement each other through an iterative process. For weakly-supervised training of SAM-Net, the semantic attribute matching loss is presented, which enables us to alleviate the photometric and geometric variations across the images simultaneously.

References

[1] K. Aberman, J. Liao, M. Shi, D. Lischinski, B. Chen, and D. Cohen-or. Neural best-buddies: Sparse cross-domain correspondence. *In: SIGGRAPH*, 2018.

[2] M. Ashikhmin. Fast texture transfer. *IEEE Comput. Graph. and Appl.*, (4):38–43, 2003.

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. ToG*, 28(3):24, 2009.

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations,. *In: ICCV*, 2009.

[5] H. Bristow, J. Valmadre, and S. Lucey. Dense semantic correspondence where every pixel is a classifier. *In: ICCV*, 2015.

[6] T. Q. Chen and M. Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv:1612.04337*, 2016.

[7] C. B. Choy, Y. Gwak, and S. Savarese. Universal correspondence network. *In: NIPS*, 2016.

[8] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. *In: SIGGRAPH*, 2001.

[9] O. Frigo, N. Sabater, J. Delon, and P. Hellier. Split and match: Example-based adaptive patch sampling for unsupervised style transfer. *In: CVPR*, 2016.

[10] L. A Gatys, A. S Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv:1508.06576*, 2015.

[11] L. A Gatys, A. S Ecker, and M. Bethge. Image style transfer using convolutional neural networks. *In: CVPR*, 2016.

[12] S. Gu, C. Chen, J. Liao, and L. Yuan. Arbitrary style transfer with deep feature reshuffle. 2018.

[13] B. Ham, M. Cho, C. Schmid, and J. Ponce. Proposal flow. *In: CVPR*, 2016.

[14] B. Ham, M. Cho, C. Schmid, and J. Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE Trans. PAMI*, 2017.

[15] K. Han, R. S. Rezende, B. Ham, K. Y. K. Wong, M. Cho, C. Schmid, and J. Ponce. Scnet: Learning semantic correspondence. *In: ICCV*, 2017.

[16] X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *In: ICCV*, 2017.

[17] Philbin. J., O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. *In: CVPR*, 2007.

[18] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. *In: NIPS*, 2015.

[19] S. Jeon, S. Kim, D. Min, and K. Sohn. Parn: Pyramidal affine regression networks for dense semantic correspondence estimation. *In: ECCV*, 2018.

[20] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song. Neural style transfer: A review. *arXiv:1705.04058*, 2017.

[21] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *In: ECCV*, 2016.

[22] J. Kim, C. Liu, F. Sha, and K. Grauman. Deformable spatial pyramid matching for fast dense correspondences. *In: CVPR*, 2013.

[23] S. Kim, S. Lin, S. Jeon, D. Min, and K. Sohn. Recurrent transformer networks for semantic correspondence. *In: NIPS*, 2018.

[24] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin, and K. Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. *In: CVPR*, 2017.

[25] S. Kim, D. Min, B. Ham, S. Lin, and K. Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. *IEEE Trans. PAMI*, 2018.

[26] S. Kim, D. Min, S. Kim, and K. Sohn. Unified confidence estimation networks for robust stereo matching. *IEEE Trans. IP*, 26(3):1299–1313, 2018.

[27] S. Kim, D. Min, S. Lin, and K. Sohn. Dctm: Discrete-continuous transformation matching for semantic flow. *In: ICCV*, 2017.

[28] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. *In: CVPR*, 2016.

[29] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. *In: ECCV*, 2016.

[30] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang. Diversified texture synthesis with feed-forward networks. *In: CVPR*, 2017.

[31] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang. Universal style transfer via feature transforms. *In: NIPS*, 2017.

[32] Y. Li, M. Liu, X. Li, M. Yang, and J. Kautz. A closed-form solution to photorealistic image stylization. *In: ECCV*, 2018.

[33] Y. Li, D. Min, M. S. Brown, M. N. Do, and J. Lu. Spm-bp: Sped-up patchmatch belief propagation for continuous mrfs. *In: ICCV*, 2015.

[34] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang. Visual attribute transfer through deep image analogy. *In: SIGGRAPH*, 2017.

[35] C. Liu, J. Yuen, and A Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. PAMI*, 33(5):815–830, 2011.

[36] J. Lu, H. Yang, D. Min, and M. N. Do. Patchmatch filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation. *In: CVPR*, 2013.

[37] M. Lu, H. Zhao, A. Yao, F. Xu, Y. Chen, and L. Zhang. Decoder network over lightweight reconstructed feature for fast semantic style transfer. *In: ICCV*, 2017.

[38] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. *CoRR, abs/1703.07511*, 2, 2017.

[39] D. Novotny, D. Larlus, and A. Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. *In: CVPR*, 2017.

[40] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Comput. Graph. and Appl.*, 21(5):34–41, 2001.

[41] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. *In: CVPR*, 2017.

[42] I. Rocco, R. Arandjelovic, and J. Sivic. End-to-end weakly-supervised semantic alignment. *In: CVPR*, 2018.

[43] I. Rocco, M. Cimpoi, R. Arandjelovic, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood consensus networks. *In: NIPS*, 2018.

[44] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *In: MICCAI*, 2015.

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *In: ICLR*, 2015.

[46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *IEEE Trans. Multimedia*, 15:1929–1958, 2014.

[47] Y. Tai, J. Jia, and C. Tang. Local color transfer via probabilistic segmentation by expectation-maximization. *In: CVPR*, 2005.

[48] T. Taniai, S. N. Sinha, and Y. Sato. Joint recovery of dense correspondence and cosegmentation in two images. *In: CVPR*, 2016.

[49] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv:1603.03417*, 2016.

[50] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. *In: CVPR*, 2017.

[51] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[52] W. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang. Style transfer via image component analysis. *IEEE Trans. Multimedia*, 15(7):1594–1601, 2013.

[53] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3d-guided cycle consistency. *In: CVPR*, 2016.

[54] T. Zhou, Y. J. Lee, S. X. Yu, and A. A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. *In: CVPR*, 2015.