General Object Pose Transformation Network from Unpaired Data

Yukun Su, Guosheng Lin†, Ruizhou Sun, Qingyao Wu†

Introduction

Limitations of the existing object transformation methods:

- Most recent approaches merely **explore human pose transformation**. Such methods require abundant **keypoints information** or **paired data**.
- In addition to human, some other **general objects** should also be able to **conduct pose transformation**, which is helpful for wider applications.
- In real life, it is **difficult** for us to **collect different postures of the same object**, which is laborious and time-costly.

Contributions

- We address the problem of general object pose transformation and **propose a unified framework with unpaired data**, which to our best knowledge, has not been well explored.
- With the proposed **four sub-blocks in the network**, we can generate **more realistic** transformed images in the desired pose preserving the original appearance and background compared to recent methods.
- Quantitative comparisons against several prior methods demonstrate the superiority of our approach, which can also be applied to **several practical applications**.

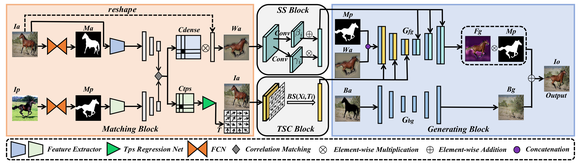
Illustrative Examples

Given the **desired pose image** (1^{st} row) and the **appearance image** (2^{nd} row) **in the absence of paired examples**, we produce the output image (3^{rd} row) in that pose and retain the appearance of object and background. We can obtain high-quality images and apply the network to different object posture modalities.

In addition, we can extend our network to complete **multi-object** and **cross-category pose transformation**. Moreover, we show the practicality of our method for several applications.

The overview of our method

- The **correlation matching block** is introduced to align the unpaired images and warp (Dense warping and TPS warping) the appearance image into the target pose.
- The **Spatial-Structural (SS) block** employs the information from the output of dense warping in the form of spatially-variant de-normalization to **progressively inject the spatial details to the generated network**.
- The **Texture-Style-Color (TSC) block** employs the information from the output of TPS warping to **preserve the appearance details when synthesizing the results**.
- The **generating block** is responsible to combine the transformed foreground object and background to **produce the output**, which is **semantically aligned to the target pose image while contextually related to the source appearance image**.



Joint training

As for dense correspondence warping, we propose to match the features of f_a and f_p by using cosine similarity as follows:

$$C_{dense} = \frac{(f_a - \mu_a)^T (f_p - \mu_p)}{||f_a - \mu_a|| ||f_p - \mu_p||}, \tag{1}$$

where μ_a and μ_p represent the mean vectors. We then calculate the weighted average to estimate the dense correspondence warping in the form as:

$$W_a = \sum \text{Softmax}(\alpha C_{dense} \cdot I_a, \text{dim} = 1), \tag{2}$$

where α is a hyper-parameter that controls the sharpness of the softmax function.

To force the network to learn a reasonable dense semantic warping, we introduce a geometric loss as follows:

$$\mathcal{L}_{geo} = ||I_p - W_a||_1. \tag{3}$$

As for TPS warping, after obtaining C_{tps} matrix like C_{dense}, then we employ a regression net to predict the corresponding control points and calculate the flow parameters T as:

$$\mathcal{L}_{tps} = ||BS(I_a, T) - I_p||_1 + \mathcal{L}_{cst}, \tag{4}$$

where BS indicates the bilinear operation. $\mathcal{L}_{cst}$ is a constraint loss.

Joint training contd.

Formally, let $F^i \in \mathbb{R}^{h \times w}$ denote the activations of the i-th layer of a deep convolutional network, we inject the dense warping information as follows:

$$\hat{F}^i = \gamma^i W_a \times \frac{F^i - \mu^i}{\sigma^i} + \beta^i W_a. \tag{5}$$

Let X^i denote the activations of the i-th layer of the network, we inject the TPS warping information as follows:

$$\hat{X}^i = \varphi^i(BS(I_a, T)) + X^i, \tag{6}$$

The final output should be semantically consistent with the desired pose image, we then minimize the semantic discrepancy between them as follow:

$$\mathcal{L}_{perc} = ||\phi_l(I_o) - \phi_l(I_p)||_2. \tag{7}$$

To encourage our network to preserve more details from source appearance image, we employ the loss as follow:

$$\mathcal{L}_{cont} = \sum_l \zeta_l[-log(\frac{1}{n_l}\sum_i \max_j A^l(\phi_l(I_o), \phi_l(I_a)))]. \tag{8}$$

In order to obtain the more realistic output, we penalize the statistic error between high-level features as follow:

$$\mathcal{L}_{style} = \sum_l ||G_l(I_o) - G_l(I_a)||_2. \tag{9}$$

To fully utilize the data under self-supervision, we construct pseudo paired data by apply random geometry transformations, we then penalize the loss as follow:

$$\mathcal{L}_{self} = \sum_l ||\phi_l(I_o) - \phi_l(I'_a)||_1. \tag{10}$$

In the matching block, since we align the image and mask from two domains, we here apply a $\mathcal{L}_1$ regularization to encourage them to be closer as follow:

$$\mathcal{L}_{reg} = ||f_a - f_p||_1. \tag{11}$$

Finally, we optimize the total loss as follow:

$$\begin{aligned}\mathcal{L}_{total} = \lambda_1\mathcal{L}_{geo} + \lambda_2\mathcal{L}_{tps} + \lambda_3\mathcal{L}_{perc} + \lambda_4\mathcal{L}_{cont} \\ + \lambda_5\mathcal{L}_{style} + \lambda_6\mathcal{L}_{self} + \lambda_7\mathcal{L}_{reg} + \lambda_8\mathcal{L}_{adv},\end{aligned} \tag{13}$$

Ablation Study

- **Experiments of each proposed functions**

Methods	Human		Mammals		Birds		Cars	
	mFID ↓	mSSIM ↑	mFID ↓	mSSIM ↑	mFID ↓	mSSIM ↑	mFID ↓	mSSIM ↑
w/o $\mathcal{L}_{perc}$	63.1	0.255	68.4	0.193	57.6	0.211	71.4	0.132
w/o $\mathcal{L}_{cont}$	52.8	0.406	58.5	0.394	49.2	0.301	62.2	0.267
w/o $\mathcal{L}_{self}$	40.6	0.601	36.6	0.521	30.7	0.528	43.1	0.452
w/o $\mathcal{L}_{reg}$	38.8	0.649	35.0	0.533	29.5	0.538	41.6	0.464
w/o Tps	38.9	0.645	35.0	0.547	29.3	0.527	41.6	0.459
w/o Dense Warp	38.7	0.651	34.8	0.554	28.9	0.539	41.3	0.464
w/o $\mathcal{L}_{style}$	39.1	0.635	35.2	0.523	29.8	0.517	41.9	0.448
TPS (TSC) ↔ Dense Warp (SS)	38.2	0.655	34.3	0.543	29.1	0.542	41.3	0.468
***Ours* (full)**	**37.6**	**0.676**	**33.9**	**0.576**	**28.3**	**0.571**	**40.8**	**0.491**

Comparisons with STOA Methods

- **Comparison between our method and the previous state-of-the-arts in *Mammals*, *Birds*, *Human and Cars* datasets.**

Methods	Mammals			Birds		
	mFID ↓	mSSIM ↑	UP ↑	mFID ↓	mSSIM ↑	UP ↑
FUNIT	78.5	0.138	4%	80.4	0.182	3%
COCO-FUNIT	70.7	0.141	7%	78.8	0.186	4%
CoCosNet	81.6	0.156	6%	64.5	0.211	6%
Ours	**33.9**	**0.576**	**83%**	**28.3**	**0.571**	**87%**

Methods	Human			Cars		
	mFID ↓	mSSIM ↑	UP ↑	mFID ↓	mSSIM ↑	UP ↑
Liquid	44.6	0.559	20%	-	-	-
Liquid++	41.4	0.567	30%	-	-	-
HoloGan	-	-	-	51.8	0.251	18%
PAGM	-	-	-	46.7	0.284	28%
Ours	**37.6**	**0.676**	**50%**	**40.8**	**0.491**	**54%**

More Application

- **Cross-category and Multi-object Pose Transformation.**

- **Video Imitation. Given a static image providing appearance and a dynamic sequential video, we can yield an unseen video of that object.**