Image-free Domain Generalization via CLIP for 3D Hand Pose Estimation
– Supplemental Document

Seongyeong Lee1,2  Hansoo Park1  Dong Uk Kim1  Jihyeon Kim1
Muhammadjon Boboev1  Seungryul Baek1

1UNIST, South Korea  2NC Soft, South Korea

In this supplemental material, we provide the details of the network architectures. Also, examples of generated text prompts are included.

1. Implementation details

Network architectures. Tables 1–2 present the details of our network architectures.

Text prompts. We presented the domain generalization framework based on the CLIP model for 3D hand pose estimation task. At this time, RGB images and text prompts are input to the CLIP model. These text prompts consist of words related to the hand domain that contain information from the hand dataset. Specifically, the sentence expresses image information in the context corresponding to the hand and the background area corresponding to the background. The composition of the sentence can be checked in Table 3, and the sentence is composed by making a combination of each element. Our text prompts consist of a total of 3,920 sentences. Some of the sentences used are specifically shown in Table 4.

### Table 1. Architecture of 2D heatmap net $f^{\text{H2D}}$

| Layer | Operation | Kernel | Dimensionality |
|-------|-----------|--------|----------------|
| -     | Input: RGB image | - | $256 \times 256 \times 3$ |
| 1     | Conv. + LeakyReLU | $3 \times 3$ | $256 \times 256 \times 64$ |
| 2     | Conv. + LeakyReLU | $3 \times 3$ | $256 \times 256 \times 64$ |
| 3     | MaxPooling | | $128 \times 128 \times 64$ |
| 4     | Conv. + LeakyReLU | $3 \times 3$ | $128 \times 128 \times 128$ |
| 5     | Conv. + LeakyReLU | $3 \times 3$ | $128 \times 128 \times 128$ |
| 6     | MaxPooling | $2 \times 2$ | $64 \times 64 \times 128$ |
| 7     | Conv. + LeakyReLU | $3 \times 3$ | $64 \times 64 \times 256$ |
| 8     | Conv. + LeakyReLU | $3 \times 3$ | $64 \times 64 \times 256$ |
| 9     | Conv. + LeakyReLU | $3 \times 3$ | $64 \times 64 \times 256$ |
| 10    | Conv. + LeakyReLU | $3 \times 3$ | $64 \times 64 \times 256$ |
| 11    | MaxPooling | $2 \times 2$ | $32 \times 32 \times 256$ |
| 12    | Conv. | $3 \times 3$ | $16 \times 16 \times 512$ |
| 13    | Conv. | $3 \times 3$ | $8 \times 8 \times 1024$ |
| 14    | MaxPooling | $8 \times 8$ | $1 \times 1 \times 1024$ |
| 15    | Flatten | | $1024$ |
| 16    | Linear | | $512$ |
| 17    | Input: (L15, L16), Concatenate | - | $1024 \times 512$ |
| 18    | Linear | | $512$ |
| 19    | Input: (L16, L18), Concatenate | - | $512 \times 512$ |
| 20    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 21    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 22    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 23    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 24    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 25    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 26    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 27    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 28    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 29    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 149$ |
| 30    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 31    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 32    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 33    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 34    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 35    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 36    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 37    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 21$ |
| 38    | Input: (L26, L28), Concatenate | - | $32 \times 32 \times 149$ |
| 39    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 40    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 41    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 42    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 43    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 44    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 128$ |
| 45    | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 21$ |
Table 2. Architecture of Poseprior net $f_{PP}$

| Layer | Operation | Kernel | Dimensionality |
|-------|-----------|--------|----------------|
| Input | 2D heatmap | - | $32 \times 32 \times 21$ |
| 1     | Conv. + LeakyReLU | $3 \times 3$ | $32 \times 32 \times 21$ |
| 2     | Conv. + LeakyReLU | $3 \times 3$ | $16 \times 16 \times 32$ |
| 3     | Conv. + LeakyReLU | $3 \times 3$ | $8 \times 8 \times 64$ |
| 4     | Conv. + LeakyReLU | $3 \times 3$ | $8 \times 8 \times 128$ |
| 5     | Conv. + LeakyReLU | $3 \times 3$ | $4 \times 4 \times 128$ |
| 6     | Flatten. | - | 2048 |
| 7     | Linear. + LeakyReLU | - | 512 |
| 8     | Linear. + LeakyReLU | - | 512 |
| 9     | Linear. | - | 63 |
| 10    | Reshape. | - | $21 \times 3$ |

Table 3. Composition of text prompts

| head | hand color | hand color | color | background |
|------|------------|------------|-------|-----------|
| a cropped image of | white | hand with | mountain | lake |
| a picture of | dark brown | right hand | bright | dark |
| one | peach | green | purple |
| | pale yellow | sky blue | black |
| a photo of | light beige | orange | red |
| a photo of | | blue | yellow |
| of right | | gray | beige |
| a cropped photo of | | pink | brown |
| a picture of | | dotted | flower |

Table 4. Some of the sentences with text prompts. Our text prompts consist of a total of 3,920 sentences. The thirty-five example sentences here are part of a total of 3,920 sentences.

- 'a picture of peach hand with sky blue background'
- 'one white hand with orange room'
- 'a photo of right black hand with dark room'
- 'a photo of right white right hand with black background'
- 'a photo of black hand with lake background'
- 'a cropped photo of peach right hand with black background'
- 'a cropped photo of white right hand with yellow room'
- 'a cropped photo of light beige right hand with blue background'
- 'a cropped photo of pale yellow hand with bright background'
- 'a cropped photo of pale yellow hand with brown room'
- 'a cropped image of right hand with yellow background'
- 'a cropped image of right hand with white background'
- 'one white hand with bright room'
- 'one peach hand with green background'
- 'a photo of peach hand with black room'
- 'a photo of right hand with green background'
- 'a cropped photo of light brown hand with brown background'
- 'a picture of right dark brown hand with blue background'
- 'a cropped image of light beige right hand with blue room'
- 'a cropped image of light beige right hand with red background'
- 'a picture of right brown hand with green background'
- 'a cropped photo of pale yellow hand with lake room'
- 'a cropped photo of light beige right hand with blue room'
- 'a cropped image of peach hand with blue room'
- 'one white hand with gray background'
- 'a cropped image of white hand with black background'
- 'a cropped image of black hand with beige background'
- 'one peach hand with mountain room'
- 'a picture of right hand with white background'
- 'a cropped photo of black hand with purple room'
- 'a picture of black hand with yellow room'
- 'a cropped photo of light beige right hand with sky blue background'
- 'a cropped photo of light yellow hand with orange background'
- 'a cropped photo of light yellow hand with yellow background'