This CVPR 2021 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Face Parsing from RGB and Depth Using Cross-Domain Mutual Learning

Jihyun Lee
KAIST
jyun.lee@kaist.ac.kr

Binod Bhattarai
Imperial College London
b.bhattarai@imperial.ac.uk

Tae-Kyun Kim
KAIST, Imperial College London
tk.kim@imperial.ac.uk

Abstract

Existing methods of face parsing have proven effective at classifying each pixel of an RGB image into different facial components. However, there is a lack of face parsing research that utilizes depth domain. To the best of our knowledge, we present the first study to exploit 2.5D data for face parsing. We introduce a novel framework to jointly learn (1) RGB face parsing, (2) depth face parsing and (3) RGB-to-depth domain translation, which can be effective even when only a small amount of annotated depth data is available for training. To this end, we also create the first RGB-D face parsing benchmarks based on CelebAMask-HQ, LaPa and Helen by utilizing an off-the-shelf 3D head reconstruction model. Overall, our approach makes two main contributions. First, our method leverages mutual learning between RGB and depth face parsing, which enables bidirectional knowledge distillation between the two data domains. Second, our method utilizes end-to-end learning of RGB-to-depth domain translation and depth face parsing, which can help overcome the scarcity of annotated depth data. We perform extensive experiments to validate the effectiveness of our method, in which we achieve state-of-the-art results in RGB face parsing. As far as we know, we also report the first results on face parsing from depth data. All experiments are conducted on our new RGB-D face parsing datasets, which are publicly available at https://github.com/jyunlee/CelebAMask-HQ-D_LaPa-D_Helen-D.

1. Introduction

Face parsing is an important research problem. It is useful in several high-level applications, including face understanding, synthesis and animation [16, 44, 45]. Recently, deep learning-based methods [14, 18, 20, 21, 32, 39, 48, 49] have proven effective for face parsing from RGB images, and a number of datasets [12, 16, 21, 29] have been published to this end. However, RGB images are sensitive to differences in lighting conditions, which makes it difficult for them to capture useful information in very bright or dark scenarios. Also, RGB images do not directly capture geometric information, which has been shown to benefit semantic segmentation [6, 8, 9, 17, 22, 24, 26]. For these reasons, several studies have considered utilizing depth or event modalities in other computer vision domains (e.g., semantic segmentation) [1, 6, 31], while no such research has yet been carried out for face parsing.

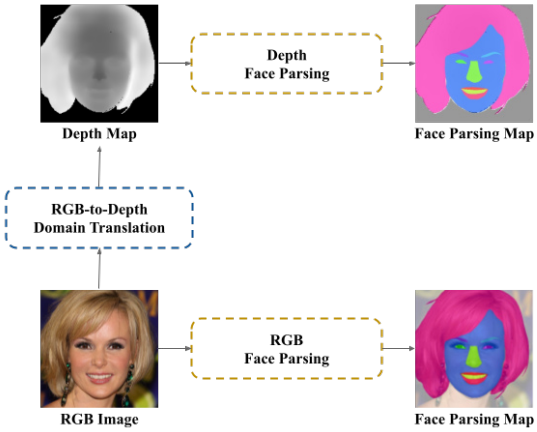


Figure 1: Our results of RGB face parsing, depth face parsing and RGB-to-depth domain translation. The proposed method can jointly learn the three tasks in an end-to-end manner.

In this work, we explore depth face parsing along with RGB face parsing in a symbiotic manner. As depth maps are significantly less sensitive to variation in illumination than conventional RGB images, depth face parsing can offer advantages for indoor applications in difficult lighting environments. Once face parsing is learned from depth data, it can also be utilized as privileged knowledge to improve the results of the conventional RGB face parsing methods [4, 5, 11, 25, 27, 43].

One of the main reasons for the comparative scarcity

of similar studies on face parsing is the lack of annotated datasets consisting of samples other than RGB images. To address this issue, we create and publish the *first* RGB-D datasets with annotated label maps, which are manufactured by utilizing an off-the-shelf 3D reconstruction model [41] on the three existing challenging RGB-based datasets: CelebAMask-HQ [16], LaPa [21] and Helen [20]. However, although we have constructed large-scale annotated depth datasets, this type of data still remains relatively scarce compared to annotated RGB data in the overall face parsing literature. To potentially utilize all existing RGB-based datasets to the fullest, we additionally introduce an RGB-to-depth domain translation network based on Pix2Pix [13] architecture in our framework to generate more depth training examples from the annotated RGB examples on the fly. In Section 3.2, we will discuss how the optimization of the RGB-to-depth domain translation network can also benefit from the end-to-end learning mechanism put forth in our framework.

Overall, our unified framework aims to jointly train (1) an RGB-based face parsing network, (2) a depth-based face parsing network and (3) an RGB-to-depth domain translation network. We first pre-train each of the base networks separately, then jointly train them in an end-to-end manner. In particular, we utilize the annotated RGB datasets not only to learn RGB face parsing, but also depth face parsing. First, we propagate a given RGB image to both the RGB face parsing network and the RGB-to-depth domain translation network. Then, we feed-forward the translated depth map to the depth face parsing network. Next, the two parallel face parsing networks perform mutual learning [46], in which activations in the last layer of the two networks are stimulated to mimic each other. Notably, most of the existing studies on domain transfer learning involve one-way transfer of supervision between different data domains [2, 25, 43], considering one as the target domain and another as the privileged domain. However, we treat both the RGB and depth domains as the target domain as well as the privileged domain to one another by allowing knowledge distillation in a *bidirectional* manner. To the best of our knowledge, this is also the first study to leverage mutual learning [46] mechanism for domain transfer learning.

Our main contributions can be summarized as follows.

- We propose a novel end-to-end method to jointly learn RGB face parsing, depth face parsing and RGB-to-depth domain translation. Our experiments demonstrate that the method enhances the performance of all three tasks. In particular, we achieve state-of-the-art results in RGB face parsing.

- To the best of our knowledge, we present the first study on face parsing from depth maps.

- As far as we know, we construct and publish the first RGB-D face datasets with semantic label maps. We estimate depth information of the CelebAMask-HQ [16], LaPa [21] and Helen [20] benchmarks by utilizing an off-the-shelf 3D head reconstruction model [41].

2. Related Work

2.1. RGB Face Parsing

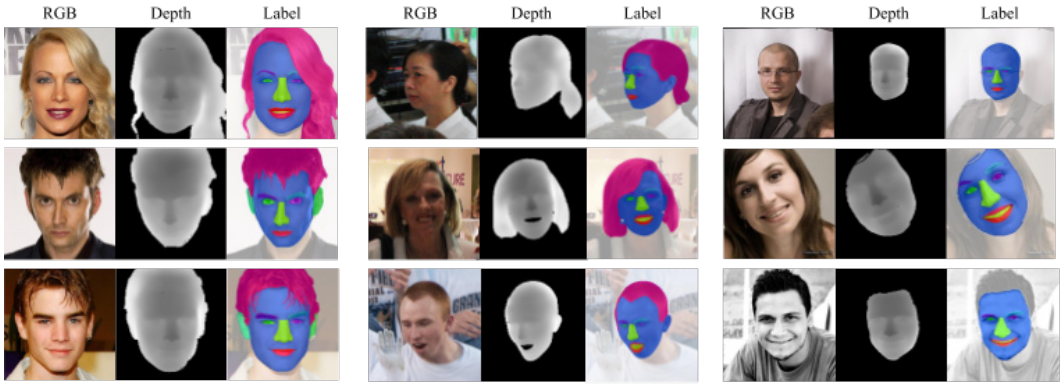
For many years, methods for face parsing from RGB images have been actively investigated. Traditional approaches learn correlation between facial regions using hand-crafted features and machine learning models, such as Gaussian Radial Basis Function (RBF) [29] and epitome model [38]. With the development of deep learning, most of recent approaches adapt deep convolutional neural networks (CNNs) for more effective feature learning [14, 18, 20, 21, 32, 39, 48, 49]. For example, Lin *et al*. introduce a CNN-based framework with a RoI Tanh-Warping operator to combine central and peripheral information [18]. Te *et al*. adapt edge-aware graph representation learning to effectively model the relation between facial components [32]. Although there exist a considerable number of studies on face parsing from RGB images, there is a lack of face parsing studies that utilize other data domains, due to the unavailability of annotated datasets.

2.2. RGB-D Semantic Segmentation

Unlike in face parsing, many researches in semantic segmentation for indoor scene understanding have been utilizing both RGB and depth data, as numerous RGB-D datasets [7, 15, 28, 30, 40] are available to this end. Compared to conventional RGB images, RGB-D images are widely known to improve the quality of semantic labeling, as depth provides additional geometric and illumination-independent features. Most previous methods fuse RGB and depth channels based on early, middle or late fusion mechanism [6, 9, 17, 22, 24]. Other works feed RGB image and HHA images encoded from depth into two separate networks and combine their predictions [8, 17, 26]. In this paper, we also take advantages of both RGB and depth modalities for face parsing. One difference of our method from most of the forementioned techniques is that it requires both RGB and depth data only at the training phase. During the test phase, face parsing can be performed from either one of RGB or depth domain.

2.3. Learning Using Privileged Information and Distillation

Privileged information [36, 37] and distillation [10] are techniques that enable a model to learn from other models and data representations. Learning Using Privileged Information (LUPI) is first introduced by Vapnik and Vashist



(a) CelebAMask-HQ-D (b) LaPa-D (c) Helen-D

Figure 2: Visual examples of the samples in CelebAMask-HQ-D, LaPa-D and Helen-D datasets.

[37], where privileged information denote additional information related to the original data that is available only at the training stage - not at the test stage. Knowledge distillation [10] can also be adapted as one of methods to incorporate privileged information, as suggested by Lopez-Paz *et al*. [23]. Knowledge distillation enables a student network to learn from soft predictions or internal feature representations of a teacher network, which can possibly be trained on privileged data domains. Rad *et al.* [25] and Yuan *et al*. [43] propose methods to benefit RGB pose estimation via distilling knowledge from a network trained on privileged depth modality. Asami *et al*. introduce a domain adaptation method for acoustic models based on the knowledge distillation framework [2]. Such existing methods usually consider one-way knowledge distillation between different data domains. However, our work considers RGB and depth modalities as privileged information to *each other* via mutual learning [46].

2.4. 3D Face Model Fitting

3D model fitting is one possible way to obtain complete geometry for a face. Most of existing generative 3D face models are based on 3D Morphable Face Model [3] framework, which separately represent shape and appearance with an assumption that all faces are in point-to-point correspondence. With the advent of machine learning, most of recent state-of-the-art methods utilize deep neural networks to learn a 3D Morphable Model and regression-based fitting [19, 33, 34, 35]. Recently, Xu *et al*. proposed a framework for 3D *head* reconstruction using a single image, that can additionally learn hair and ear geometry along with the face geometry [41]. In our study, we also utilize this publicly available 3D head reconstruction model [41] to acquire pseudo-ground truth depth maps that correspond to an existing RGB face dataset with semantic labels.

3. Proposed Method

In Section 3.1, we first discuss the construction of the *first* RGB-D face datasets with semantic label maps. Then, in Section 3.2, we explain the proposed framework that can jointly learn RGB face parsing, depth face parsing network and RGB-to-depth domain translation.

3.1. First RGB-D Datasets for Face Parsing

In the current face parsing literature, a number of annotated RGB datasets are available while no such depth dataset exists. To address this issue, we construct new pseudo-ground truth RGB-D datasets with semantic annotations by following the method proposed by Bodur et al. [4]. Two differences of our dataset construction from [4] is that (1) our dataset contains pixel-level labels for segmentation, rather than image-level labels for classification, and (2) we utilize a recently published 3D head reconstruction model [41], which allows estimation for additional hair and ear regions. By utilizing this off-the-shelf method [41], we first reconstruct 3D head meshes from the RGB images in CelebAMask-HQ [16], LaPa [21] and Helen [20] datasets. Then, we project the obtained meshes to acquire depth maps that correspond to the original RGB image and label map pairs. However, we have discarded very few samples with low reconstruction quality. Table 1 shows the resulting number of depth-augmented samples that we have obtained from the three original datasets.

For CelebAMask-HQ dataset, we have also adjusted the labels of classes that the reconstruction model does not render and have obtained the label maps of *13 classes* as a result: *background*, *skin*, *left eyebrow*, *right eyebrow*, *left eye*, *right eye*, *left ear*, *right ear*, *nose*, *mouth*, *upper lip*, *lower lip* and *hair*. For Helen dataset, we have applied an additional post-processing step to discard depth and label information for hair regions. It is widely known that Helen has inaccurate annotations for hair [18, 21]. Since the 3D

reconstruction model [41] produces a hair mesh based on a hair segmentation map, applying such post-processing has helped us to obtain more plausible results. For future use, we denote the depth-augmented version of CelebAMask-HQ, LaPa and Helen as CelebAMask-HQ-D, LaPa-D and Helen-D, respectively. Figure 2 shows the visual examples of samples in the constructed datasets.

Dataset	Train set	Validation set	Test set	Total
CelebAMask-HQ [16]	24,183	2,993	2,284	30,000
CelebAMask-HQ-D	24,161	2,982	2,821	29,964
LaPa [21]	18,168	2,000	2,000	21,505
LaPa-D	17,656	1,934	1,915	21,428
Helen [20]	2,000	230	100	2,330
Helen-D	1,897	216	95	2,208

Table 1: Number of samples in train, validation and test sets of CelebAMask-HQ-D, LaPa-D and Helen-D.

3.2. End-to-End Face Parsing From RGB and Depth

Our framework consists of three base models: an RGB-based face parsing network, a depth-based face parsing network and an RGB-to-depth domain translation network. The overall pipeline of our method is shown in Figure 3. We first explain each of the base models separately. Afterwards, we introduce our novel learning mechanism to jointly optimize the three models in an end-to-end manner.

RGB Face Parsing Network. The RGB face parsing network aims to classify each pixel of an input *RGB image* into different facial components (see the bottom horizontal branch in Figure 3). We adopt a deep CNN-based architecture to instantiate the network, as it has shown state-of-the-art results in the recent face parsing literatures [18, 20, 32, 49]. However, we would like to note that our framework is generic, and it is not constrained to one base model architecture. The original loss function to optimize the parameters of the RGB face parsing network, which we denote by $\mathcal{L}_{CLS}^{RGB}$, can also vary depending on your engineering choice for the base network architecture. One common scenario would be to use cross entropy loss, since it is the most widely adapted for training face parsing networks.

Depth Face Parsing Network. The depth face parsing network aims to classify each pixel of an input *depth map* into different semantic components (see the top horizontal branch in Figure 3). To instantiate the depth face parsing network, it would be a natural choice to adapt the same network architecture and loss function that are used for the RGB face parsing network. For future convenience, we denote the original loss function for the depth face parsing network as $\mathcal{L}_{CLS}^{Depth}$.

RGB-to-Depth Translation Network. The RGB-to-depth translation network aims to convert an input RGB face image into the corresponding depth map (see the left vertical branch in Figure 3). For the network architecture, we adopt a conditional generative adversarial network (GAN), which has been proven to be effective in various image-to-image translation tasks [13, 50]. We denote the original loss function for the RGB-to-depth domain translation network as $\mathcal{L}_{DT}$, for which we can utilize the one suggested in [13]:

$$\mathcal{L}_{DT} = \mathcal{L}_{CGAN} + \lambda \mathcal{L}_{L1}, \tag{1}$$

where $\mathcal{L}_{CGAN}$ and $\mathcal{L}_{L1}$ are conditional GAN loss and L1 loss, respectively. λ is a hyper-parameter to control the weight of the L1 loss term. The comprehensive definitions of $\mathcal{L}_{CGAN}$ and $\mathcal{L}_{L1}$ can be written as follows:

$$\begin{aligned}\mathcal{L}_{CGAN} = \mathbb{E}_{x_R,x_D}[\log D(x_R, x_D)] + \\ \mathbb{E}_{x_R,z}[\log(1 - D(x_R, G(x_R, z))],\end{aligned} \tag{2}$$

$$\mathcal{L}_{L1} = \mathbb{E}_{x_R,x_D,z}[\|x_D - G(x_R, z)\|_1], \tag{3}$$

where G and D indicate a generator network and a discriminator network, respectively. x_R represents an RGB image and x_D denotes its ground-truth depth counterpart. z represents a noise vector sampled from a Gaussian distribution.

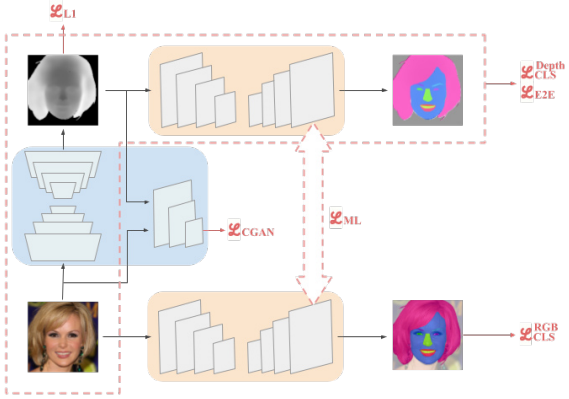


Figure 3: Detailed overview of our method. Besides loss functions to independently train each of the base models (i.e., $\mathcal{L}_{CLS}^{RGB}$, $\mathcal{L}_{CLS}^{Depth}$, $\mathcal{L}_{CGAN}$ and $\mathcal{L}_{L1}$), we propose to utilize additional loss terms to stimulate joint learning between them: $\mathcal{L}_{ML}$ and $\mathcal{L}_{E2E}$. $\mathcal{L}_{ML}$ is adapted for mutual learning between RGB and depth face parsing networks (dashed arrow) and $\mathcal{L}_{E2E}$ is used for end-to-end learning of an RGB-to-depth domain translation network and a depth face parsing network (dashed box).

Joint Learning Mechanism. For training our framework, we first separately pre-train each of the base networks using the constructed RGB-D dataset (from Section 3.1), then jointly train them in an end-to-end manner. To this end, we

introduce a novel loss formula to stimulate joint learning between the three base models:

$$\begin{aligned}\mathcal{L} = \lambda_1 \, \mathbb{E}_{x_R, x_D} \, \| F_R(x_R) - F_D(x_D) \|_2^2 \, + \\ \lambda_2 \, \mathbb{E}_{x_R, y_R, z} \texttt{FPLoss}(F_D(G(x_R, z)), y_R) + \\ \lambda_3 \, \mathcal{L}_{CLS}^{RGB} + \lambda_4 \, \mathcal{L}_{CLS}^{Depth} + \lambda_5 \, \mathcal{L}_{DT},\end{aligned} \quad (4)$$

where F_R and F_D denote the RGB face parsing network and the depth face parsing network, respectively. $\texttt{FPLoss}(\cdot, \cdot)$ denotes a loss function that measures the deviation between a prediction map of the face parsing network and a label map, for which we can adapt the same loss formula used for $\mathcal{L}_{CLS}^{RGB}$ and $\mathcal{L}_{CLS}^{Depth}$ (e.g., cross entropy loss). y_R represents the label map corresponding to x_R. Lastly, λ_1, λ_2, λ_3, λ_4 and λ_5 are hyper-parameters to control weights between the loss terms. Overall, the training objective of all the networks involved in our framework is to minimize $\mathcal{L}$, except for the discriminator of the RGB-to-depth translation network which competes to maximize it.

Please note that the first and the second loss terms in Equation 4 are newly introduced for the joint training phase. The first loss term, which we denote by $\mathcal{L}_{ML}$, is used to stimulate mutual learning [46] between RGB-based and depth-based face parsing networks. It encourages activations in the last layer of the two networks to mimic each other, which enables *bidirectional* knowledge distillation between RGB and depth domains. We would like to emphasize that, unlike most of existing methods for domain transfer learning that involve one-way transfer of supervision [2, 10, 25, 43], we propose to utilize such distillation loss for optimizing *both* RGB-based and depth-based networks. In this way, we can fully utilize $\mathcal{L}_{ML}$ to benefit the learning of both RGB face parsing and depth face parsing simultaneously. To the best of our knowledge, this is the first study to leverage mutual learning [46] mechanism for domain transfer learning, which we term *cross-domain mutual learning*.

The second loss term, which we denote by $\mathcal{L}_{E2E}$, allows end-to-end learning of the depth face parsing network and the RGB-to-depth domain translation network. It enables the depth face parsing network to be trained on data generated by the RGB-to-depth translation network. We would like to note that learning from $\mathcal{L}_{E2E}$ does not require any annotated *depth* data. Instead, it utilizes annotated *RGB* data (i.e., x_R and y_R pairs) to learn depth face parsing (see Equation 4). Considering the current face parsing literature where only RGB datasets with semantic annotations are abundant, utilizing this loss term can effectively help overcome the scarcity of depth data in learning depth face parsing. From another point of view, it can also benefit the learning of RGB-to-depth translation by incorporating additional supervision from the auxiliary face parsing task.

4. Experimental Validation

In this section, we conduct experiments to investigate the effectiveness of our method in learning RGB face parsing, depth face parsing and RGB-to-depth domain translation.

4.1. Dataset and Evaluation Metric

Dataset. Our experiments are conducted on CelebAMask-HQ-D, LaPa-D and Helen-D datasets, whose construction process has been described in Section 3.1. We divide each dataset into train, validation and test sets as shown in Table 1 by following the split protocol used in CelebAMask-HQ, LaPa and Helen for CelebAMask-HQ-D, LaPa-D and Helen-D, respectively. However, we utilize only 10% of the depth and label map pairs in each dataset for training our framework. Our intention is to show that the proposed method can be effective even in a situation where the annotated depth data is not sufficiently available, which is more practical scenario considering the current face parsing literature.

Evaluation Metrics. For RGB and depth face parsing, we evaluate our results on the standard metric for face parsing: F1-score. On CelebAMask-HQ-D and LaPa-D datasets, we report F1-scores corresponding to each category, excluding the background, and their mean F1-score. On Helen-D dataset, we employ the overall F1-score over the merged brows, eyes, nose and mouth categories, in order to maintain consistency with the previous works [18, 20, 21, 32, 48]. For RGB-to-depth domain translation, we evaluate our results in both quantitative and qualitative manners. For the quantitative metric, we employ L1 distance between predicted depth maps and their corresponding ground truth depth maps.

4.2. Implementation Details

Network Architectures. We adapt the same network design for both RGB and depth face parsing networks. Throughout the evaluation, we consider two architectural options for them: EAGRNet [32] and BiSeNet [42]. EAGRNet is a recently proposed method for face parsing and has shown state-of-the-art results on CelebAMask-HQ, LaPa and Helen datasets. It utilizes edge-aware graph representation learning to effectively model the relation between different facial regions (see [32] for more architectural details). BiSeNet is a CNN-based model that incorporates a *Spatial Path* and a *Context Path*, which is originally proposed for general semantic segmentation. However, its public implementation tuned for face parsing [1] has shown to be effective, and it has become one of the most popular repositories in face parsing domain. For this reason, we also utilize this implementation of BiSeNet for evaluating our method. For

[1] https://github.com/zllrunning/face-parsing.PyTorch

Method	Skin	Nose	L-Eye	R-Eye	L-Brow	R-Brow	L-Ear	R-Ear	Mouth	U-Lip	L-Lip	Hair	Mean
H. Zhao *et al.* [47]	94.8	90.3	79.9	80.1	77.3	78.0	75.6	73.1	89.8	87.1	88.8	90.4	83.8
C. Lee *et al.* [16]	95.5	85.6	84.3	85.2	81.4	81.2	84.9	83.1	63.4	88.9	90.1	86.6	84.2
EAGRNet$_{\text{RGB}}$ [32]	96.0	93.2	87.6	87.8	85.6	85.6	86.2	81.9	91.4	88.3	91.0	95.1	89.1
EAGRNet$_{\text{RGB}}$ + our method	96.3	93.8	88.9	88.7	86.1	86.0	88.3	86.6	92.4	89.6	91.0	95.4	**90.3**
BiSeNet$_{\text{RGB}}$ [42]	96.6	94.0	71.1	70.1	66.9	67.4	68.4	65.7	86.3	88.5	90.6	95.4	80.1
BiSeNet$_{\text{RGB}}$ + our method	96.5	94.0	88.7	88.7	84.0	84.1	80.5	80.8	86.6	88.7	90.8	95.4	**88.2**

Table 2: Comparison of RGB face parsing results on CelebAMask-HQ (in F1-score). EAGRNet$_{\text{RGB}}$ and BiSeNet$_{\text{RGB}}$ represent EAGRNet and BiSeNet baselines trained on RGB data, respectively.

the RGB-to-depth domain translation network, we adapt the design of Pix2Pix [13], composed of an U-Net generator and a PixelGAN discriminator.

Training Details. During the pre-training phase, we separately train each of three base models from scratch. For RGB and depth face parsing networks, we use stochastic gradient descent with an initial learning rate of $1e-3$ for EAGRnet and $1e-2$ for BiSeNet, along with a polynomial learning rate decay schedule. We train them for $100K$ iterations using a batch size of 8, with momentum and weight decay parameters fixed to 0.9 and $5e-4$, respectively. We would also like to note that the loss formula originally proposed for the baseline face parsing network (i.e., EAGRNet or BiSeNet) is adapted for $\mathcal{L}_{CLS}^{RGB}$, $\mathcal{L}_{CLS}^{Depth}$ and $\mathcal{L}_{E2E}$ (see [32, 42] for more details). For Pix2Pix, we use an Adam optimizer with an initial learning rate of $1e-3$ and a polynomial learning rate decay schedule. We train it for $200K$ iterations using a batch size of 1. The hyper-parameter λ Equation 1 is set to 100, as in [13]. During the joint training phase, the optimization techniques for each network are applied in the same way, except for the initial learning rate which is reduced to one-tenth of its original value. Also, in order to simultaneously incorporate $\mathcal{L}_{CLS}^{Depth}$ and $\mathcal{L}_{E2E}$, the depth face parsing network is trained on mini-batches, where each of them is composed of half of ground truth depth maps and half of depth maps generated by the RGB-to-depth domain translation network on the fly. The hyper-parameters in Equation 4, λ_1, λ_2, λ_3, λ_4 and λ_5, are set to 1, 1, 0.5, 0.5 and 1, respectively.

4.3. Experimental Analysis and Comparison

We report the experimental results of our method in learning RGB face parsing, depth face parsing and RGB-to-depth domain translation. The experiments are mainly conducted on CelebAMask-HQ-D dataset (see 4.3.1), while we also include the results on LaPa-D and Helen-D datasets (see 4.3.2).

4.3.1 Evaluation on CelebAMask-HQ-D Dataset

RGB Face Parsing. We first conduct experiments to compare our method with the baselines [32, 42] and the existing works [16, 47] in RGB face parsing on CelebAMask-HQ-D dataset. Table 2 shows the quantitative results in F1-score. Our method leads to an improvement over the EAGRNet baseline by 1.2 and the BiSeNet baseline by 8.1 in mean F1-score. We would like to emphasize that the baseline EAGRNet is the current state-of-the-art work on RGB face parsing, and our method is shown to further improve its performance by incorporating privileged knowledge from depth. In Figure 4, we also provide a few qualitative examples of the experimental results. We observe that our method yields more accurate parsing results compared to the baseline, especially for eye and brow regions (see dashed boxes in Figure 4).

Figure 4: Qualitative comparison of RGB face parsing results on CelebAMask-HQ-D (best viewed in color).

Depth Face Parsing. To the best of our knowledge, we introduce the first results on face parsing from depth maps. The quantitative results in F1-score are shown in Table 3. We observe that our method improves the EAGRNet and BiSeNet baseline results by incorporating joint learning losses (i.e., $\mathcal{L}_{E2E}$ and $\mathcal{L}_{ML}$). For your reference, we also include a number of visual examples of our results in Figure 5. Compared to the baseline, our method yields more reliable parsing outcomes for eye, brow and hair regions (see dashed boxes in Figure 5). We would also like to note that, although RGB images capture richer texture information than depth maps, the performance gap between RGB

Method	Skin	Nose	L-Eye	R-Eye	L-Brow	R-Brow	L-Ear	R-Ear	I-Mouth	U-Lip	L-Lip	Hair	Mean
$\text{EAGRNet}_{\text{Depth}}$	93.9	89.8	78.0	77.6	72.5	72.7	73.7	70.9	80.8	76.9	82.5	96.2	80.5
$\text{EAGRNet}_{\text{Depth}}$ + our method	94.6	91.7	79.1	79.5	73.8	73.8	74.4	71.8	81.2	77.3	82.4	96.5	**81.3**
$\text{BiSeNet}_{\text{Depth}}$	94.3	92.2	78.0	75.5	70.1	70.1	56.3	54.9	63.1	73.9	82.4	94.7	75.5
$\text{BiSeNet}_{\text{Depth}}$ + our method	94.9	92.2	81.5	81.1	72.4	72.7	65.3	58.4	67.7	78.0	83.6	95.9	**78.6**

Table 3: Comparison of depth face parsing results on CelebAMask-HQ-D (in F1-score). $\text{EAGRNet}_{\text{Depth}}$ and $\text{BiSeNet}_{\text{Depth}}$ denote EAGRNet and BiSeNet baselines trained on depth data, respectively.

face parsing and depth face parsing are shown not to be significant (see Table 2 and 3). This suggests that depth face parsing can be an effective alternative to conventional RGB face parsing for selective environments (e.g., very bright or dark scenarios).

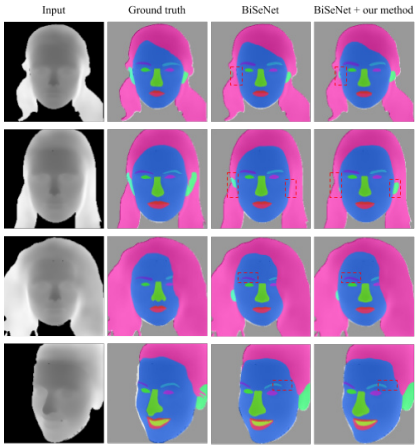


Figure 5: Qualitative comparison of depth face parsing results on CelebAMask-HQ-D (best viewed in color).

RGB-to-Depth Domain Translation. We compare RGB-to-depth domain translation results of our method to those of the Pix2Pix baseline. Evaluations are performed on CelebAMask-HQ-D dataset in both quantitative and qualitative manners. Table 4 demonstrates the quantitative results in average L1 distance between the predicted depth maps and the corresponding ground truth depth maps in the test set. Our method leads to an improvement over the Pix2Pix baseline by approximately 20% in L1 distance by incorporating an auxiliary face parsing loss (i.e., $\mathcal{L}_{E2E}$) from joint learning. We also report a number of qualitative examples of our RGB-to-depth domain translation results in Figure 6. As shown in the top two rows, our method yields more detailed depth prediction for regions around eyes and nose compared to the baseline. The bottom two rows demonstrate that our method can more precisely estimate for boundary regions between forehead and hair (see dashed boxes in Figure 6).

Method	Pix2Pix [13]	Pix2Pix + our method
L1 distance	0.122	0.099

Table 4: Comparison of RGB-to-depth domain translation results on CelebAMask-HQ-D.

Figure 6: Qualitative comparison of RGB-to-depth domain translation results on CelebAMask-HQ-D.

4.3.2 Evaluation on LaPa-D and Helen-D Datasets

We also conduct experiments to evaluate the effectiveness of our method on LaPa-D and Helen-D datasets. Table 5 and 6 demonstrate the quantitative results on RGB face parsing and depth face parsing in F1-score. Similar to the results in Section 4.3.1 on CelebAMask-HQ-D dataset, our method leads to an improvement over the baseline in both RGB face parsing and depth face parsing. However, unlike on CelebAMask-HQ and LaPa datasets, the state-of-the-art *baseline* RGB face parsing results on Helen dataset reported in the original EAGRNet paper [32] could not be reproduced. However, we would like to emphasize that our framework is not specific to a particular baseline face parsing method, and it demonstrates consistent performance gains compared to the baselines.

4.4. Ablation Study

Different Loss Components. We explore the effectiveness of different loss components on learning RGB face parsing

Method	Skin	Hair	L-eye	R-eye	U-lip	I-mouth	L-lip	Nose	L-brow	R-brow	Mean
EAGRNet $_{\text{RGB}}$ [32]	97.3	96.2	89.5	90.0	88.1	90.0	89.0	97.1	86.5	87.0	91.1
EAGRNet $_{\text{RGB}}$ + our method	97.3	95.7	90.9	91.0	87.7	89.8	90.0	97.3	88.8	89.1	**91.9**
EAGRNet $_{\text{Depth}}$	92.5	94.6	74.6	74.6	72.6	80.6	77.6	93.1	73.0	72.7	80.7
EAGRNet $_{\text{Depth}}$ + our method	92.5	94.6	75.7	75.6	72.4	80.7	77.1	93.0	74.2	74.0	**81.1**

Table 5: Comparison of RGB face parsing and depth face parsing results of EAGRNet on LaPa-D dataset (in F1-score).

Method	Skin	Nose	U-lip	I-mouth	L-lip	Eyes	Brows	Mouth	Overall
EAGRNet $_{\text{RGB}}$ [32]	91.8	94.4	72.9	83.7	85.8	87.8	80.5	92.5	90.6
EAGRNet $_{\text{RGB}}$ + our method	94.3	94.6	73.9	83.7	86.2	87.7	81.6	92.7	**91.0**
EAGRNet $_{\text{Depth}}$	95.4	91.3	65.7	72.2	76.3	73.5	65.3	87.5	83.7
EAGRNet $_{\text{Depth}}$ + our method	95.5	91.3	65.8	77.4	75.7	73.7	67.1	89.2	**84.5**

Table 6: Comparison of RGB face parsing and depth face parsing results of EAGRNet on Helen-D dataset (in F1-score).

and depth face parsing. Specifically, we consider different combinations of the loss terms that are additionally introduced during the joint training phase: $\mathcal{L}_{E2E}$ and $\mathcal{L}_{ML}$. The experimental results in class mean F1-score are provided in Table 7. We observe that each of the two loss terms leads to an improvement over the baseline, while the best results can be achieved when both terms are utilized.

Method	Mean F1-score
EAGRNet $_{\text{RGB}}$ [32]	89.1
EAGRNet $_{\text{RGB}}$ + $\mathcal{L}_{ML}$	90.2
EAGRNet $_{\text{RGB}}$ + $\mathcal{L}_{E2E}$	N/A
EAGRNet $_{\text{RGB}}$ + $\mathcal{L}_{ML}$ + $\mathcal{L}_{E2E}$	**90.3**
EAGRNet $_{\text{Depth}}$	80.5
EAGRNet $_{\text{Depth}}$ + $\mathcal{L}_{ML}$	81.2
EAGRNet $_{\text{Depth}}$ + $\mathcal{L}_{E2E}$	80.8
EAGRNet $_{\text{Depth}}$ + $\mathcal{L}_{ML}$ + $\mathcal{L}_{E2E}$	**81.3**

Table 7: Comparison of depth face parsing results with respect to different loss components (on CelebAMask-HQ-D).

Size of Annotated Depth Dataset. We compare the face parsing results with varying the size of the annotated depth dataset available for training. In Figure 7, we report the experimental results on RGB and depth face parsing in class mean F1-score. For the dataset size, 1%, 10% and 100% indicate the fractions of depth and label map pairs of CelebAMask-HQ-D that are used to train our framework. For depth face parsing, we observe a decrease in performance when a smaller annotated dataset is available. However, the RGB face parsing results of our method are not significantly affected by the size of the annotated depth dataset, and it has led to a noticeable increase over the baseline performance in all cases. This indicates that, even in a situation where a very small amount of depth and label pairs are available, domain transfer learning from depth can effectively improve the performance of RGB face parsing.

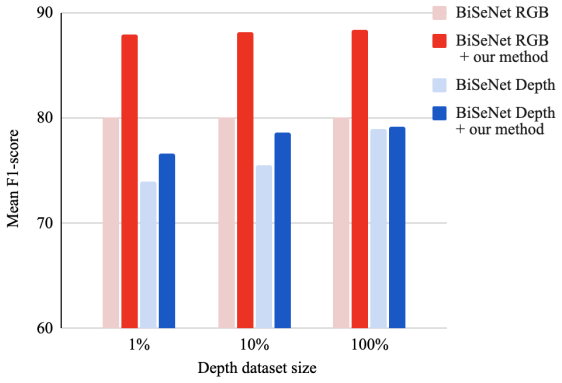


Figure 7: Comparison of face parsing results of BiSeNet with respect to different sizes of the annotated depth dataset (on CelebAMask-HQ-D).

5. Conclusions and Future Work

In this paper, we propose a novel framework to jointly learn RGB face parsing, depth face parsing and RGB-to-depth domain translation. To this end, we construct the first pseudo-ground truth RGB-D face datasets with semantic label maps. For future work, one of the remaining challenges would be to acquire a real RGB-D face dataset with semantic annotations. Although our method has achieved satisfactory results on the constructed datasets, a domain gap between the depth maps estimated by the 3D reconstruction method [41] and real depth maps may still exist. As this work has introduced some potential usefulness of an RGB-D dataset in face parsing, it would be an interesting research direction to investigate a method that can further utilize depth modality for face parsing, possibly with real RGB-D data.

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *CVPR Workshops*, 2019.

[2] Taichi Asami, Ryo Masumura, Yoshikazu Yamaguchi, Hirokazu Masataki, and Yushi Aono. Domain adaptation of dnn acoustic models using knowledge distillation. In *ICASSP*, pages 5185–5189. IEEE, 2017.

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999.

[4] Rumeysa Bodur, Binod Bhattarai, and Tae-Kyun Kim. 3d dense geometry-guided facial expression synthesis by adversarial learning. In *WACV*, 2021.

[5] Yunpeng Chen, Xiaojie Jin, Jiashi Feng, and Shuicheng Yan. Training group orthogonal neural networks with privileged information. *arXiv preprint arXiv:1701.06772*, 2017.

[6] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017.

[8] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, pages 345–360. Springer, 2014.

[9] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *ACCV*, pages 213–228. Springer, 2016.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[11] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *CVPR*, pages 826–834, 2016.

[12] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2008.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[14] Aaron S Jackson, Michel Valstar, and Georgios Tzimiropoulos. A cnn cascade for landmark guided semantic part segmentation. In *ECCV*, pages 143–155. Springer, 2016.

[15] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, pages 1817–1824. IEEE, 2011.

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020.

[17] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *ECCV*, pages 541–557. Springer, 2016.

[18] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *CVPR*, pages 5654–5663, 2019.

[19] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *CVPR*, pages 5891–5900, 2020.

[20] Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang. Multi-objective convolutional learning for face labeling. In *CVPR*, pages 3451–3459, 2015.

[21] Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In *AAAI*, pages 11637–11644, 2020.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[23] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[24] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *ICCV*, pages 4980–4989, 2017.

[25] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Domain transfer for 3d pose estimation from color images without manual annotations. In *ACCV*, pages 69–84. Springer, 2018.

[26] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 39(4):640–651, 2017.

[27] Zhiyuan Shi and Tae-Kyun Kim. Learning and refining of privileged information-based rnns for action recognition from depth sequences. In *CVPR*, pages 3461–3470, 2017.

[28] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012.

[29] Brandon M Smith, Li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. Exemplar-based face parsing. In *CVPR*, pages 3484–3491, 2013.

[30] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.

[31] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, pages 1746–1754, 2017.

[32] Gusi Te, Yinglu Liu, Wei Hu, Hailin Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing. In *ECCV*, pages 258–274. Springer, 2020.

[33] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, pages 2549–2559, 2018.

[34] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, pages 1126–1135, 2019.

[35] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *TPAMI*, 2019.

[36] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015.

[37] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

[38] Jonathan Warrell and Simon JD Prince. Labelfaces: Parsing facial features by multiclass labeling with an epitome prior. In *ICIP*, pages 2481–2484. IEEE, 2009.

[39] Zhen Wei, Yao Sun, Jinqiao Wang, Hanjiang Lai, and Si Liu. Learning adaptive receptive fields for deep image parsing network. In *CVPR*, pages 2434–2442, 2017.

[40] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, pages 1625–1632, 2013.

[41] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3d portrait from a single image. In *CVPR*, pages 7710–7720, 2020.

[42] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018.

[43] Shanxin Yuan, Bjorn Stenger, and Tae-Kyun Kim. 3d hand pose estimation from rgb using privileged learning with depth data. In *ICCV Workshops*, pages 0–0, 2019.

[44] He Zhang, Benjamin S Riggan, Shuowen Hu, Nathaniel J Short, and Vishal M Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *IJCV*, 127(6-7):845–862, 2019.

[45] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016.

[46] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018.

[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[48] Lei Zhou, Zhi Liu, and Xiangjian He. Face parsing via a fully-convolutional continuous crf neural network. *arXiv preprint arXiv:1708.03736*, 2017.

[49] Yisu Zhou, Xiaolin Hu, and Bo Zhang. Interlinked convolutional neural networks for face parsing. In *ISNN*, pages 222–231. Springer, 2015.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.