Teacher-Student Adversarial Depth Hallucination to Improve Face Recognition

Hardik Uppal Alireza Sepas-Moghaddam Michael Greenspan Ali Etemad
Queen's University, Canada

Abstract

We present the Teacher-Student Generative Adversarial Network (TS-GAN) to generate depth images from single RGB images in order to boost the performance of face recognition systems. For our method to generalize well across unseen datasets, we design two components in the architecture, a teacher and a student. The teacher, which itself consists of a generator and a discriminator, learns a latent mapping between input RGB and paired depth images in a supervised fashion. The student, which consists of two generators (one shared with the teacher) and a discriminator, learns from new RGB data with no available paired depth information, for improved generalization. The fully trained shared generator can then be used in runtime to hallucinate depth from RGB for downstream applications such as face recognition. We perform rigorous experiments to show the superiority of TS-GAN over other methods in generating synthetic depth images. Moreover, face recognition experiments demonstrate that our hallucinated depth along with the input RGB images boost performance across various architectures when compared to a single RGB modality by average values of +1.2%, +2.6%, and +2.6% for IIIT-D, EURECOM, and LFW datasets respectively. We make our implementation public at: https://github.com/hardik-uppal/teacher-student-gan.git.

1. Introduction

Facial recognition is an active research area, which has recently witnessed considerable progress thanks primarily to the effectiveness of deep neural networks such as AlexNet [23], VGG [38], FaceNet [34], ResNet [12] and others. RGB-based face recognition methods tend to be generally sensitive to facial and environmental variations like illumination, occlusions, and poses [35, 48, 1, 29]. Utilizing the depth information, acquired with an RGB-D sensor such as the Microsoft Kinect or Intel Realsense, alongside RGB allows models to learn more robust face representations. This is because depth provides complementary geometric information about the intrinsic shape of the face, further boosting recognition performance. Additionally, RGB-D facial recognition methods are known to be less sensitive than pure RGB approaches to pose and illumination variations [41, 3, 11, 42]. Despite these advantages, while RGB sensors are ubiquitous, depth sensors have been less prevalent, resulting in an over-reliance on RGB alone. To tackle this, we present a method that uses available paired RGB-D training data to learn to *hallucinate* (i.e. generate synthetic) depth images, even for datasets for which corresponding ground-truth depth information is absent.

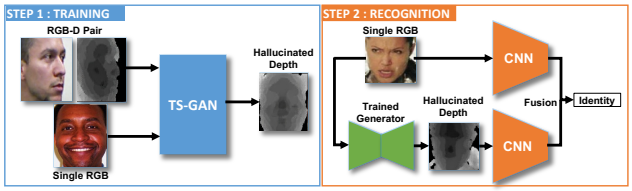

Figure 1. The proposed framework for our method. The first step (blue) trains the generator for synthesizing depth from RGB images, while the second step (orange) tests the efficacy of the synthesized depth images by using it in face recognition pipelines.

Generative Adversarial Networks (GANs) [8] and its variants (e.g., cGan [31], pix2pix [17], CycleGan [50], StackGAN [47], StyleGAN [20], etc.) have proven to be viable solutions for data synthesis in many application domains. In the context of facial images, GANs have been widely used to generate very high-quality RGB images when trained on large-scale datasets such as FFHQ [20, 21] and CelebA-HQ [19]. Nonetheless, only a limited number of past works have attempted to synthesize depth from corresponding RGB images using a conditional GAN [33], CycleGAN [24], and a Fully Convolutional Network (FCN) [4]. Although cGAN has achieved impressive results for depth synthesis using paired RGB-D sets [33], it does not easily generalize to new test examples for which paired samples are not available, especially when the images are from an entirely different dataset with drastically different poses, expressions, and occlusions. CycleGAN [50] attempts to overcome this shortcoming through unpaired training with the aim of generalizing well to new test examples. However, as stated in [50], CycleGAN does not deal well with translating geometric shapes and features.

In this work, we propose a deep architecture using a

novel Teacher-Student GAN (TS-GAN) to generate depth images from RGB images for which no corresponding depth information is available. Our end-to-end model consists of two components, a teacher and a student. The teacher consists of a fully convolutional encoder-decoder network as a generator along with a fully convolutional classification network as the discriminator. The generator takes RGB images as inputs and aims to output the corresponding depth images. In essence, our teacher aims to learn an initial latent mapping between RGB and co-registered depth images. The student consists of two generators in the form of encoder-decoders, one of which is shared with the teacher, along with a fully convolutional discriminator. The student takes as its input an RGB image for which the corresponding depth image is not available and maps it onto the depth domain as guided by the teacher. The purpose here is for the student to further refine the strict mapping learned by the teacher and allow for better generalization through a less constrained training scheme. We demonstrate the high quality of our hallucinated depth images by comparing them to ground truth depth and several state-of-the-art depth generation alternatives. The performance of our approach for using the generated depth in facial recognition is then validated for two RGB-D datasets, IIIT-D RGB-D and EURECOM KinectFaceDb, across various facial recognition networks. The results show that the depth images generated using our approach enable a performance as good as, or in some cases surprisingly even better than using the ground-truth depth originally available in the dataset, and that it gives a significant boost to recognition accuracy as compared to a pure RGB facial recognition system. We also evaluate the performance of our approach for an in-the-wild RGB dataset, Labeled-Faces-in-Wild (LFW), where no depth information is originally available, and show that the addition of hallucinated depth by our proposed method can considerably boost the recognition results by +2.4% with SE-ResNet-50 architecture.

Our contributions are summarized as follows. (**1**) A novel teacher-student adversarial architecture is proposed to generate realistic depth images from a single RGB image. Our method uses a student architecture to refine the strict latent mapping between RGB and D domains learned by the teacher to obtain a more generalizable and less constrained relationship. (**2**) Our assessments reveal that our method creates realistic synthetic depth images as compared to the original co-registered depth images (where available) and other techniques. We then utilize the synthetic depth for RGB-D facial recognition and show that multimodal solutions that utilize the depth images produced by our method perform as good as using the ground-truth depths. We also show that the facial recognition performance increases when utilizing our method to generate depth for an RGB-only dataset and subsequently combining the generated depth and original RGB images in a multimodal network. (3) We make our implementation publicly available[1] to enable reproducibility and future comparisons.

2. Related Work

2.1. Depth Generation from RGB Images

A number of methods have been proposed to estimate depth information from other modalities such as stereo vision [7, 6, 2] and multi-view images [44]. Here, given our goal in this paper, we only review methods that generate depth images from RGB data.

The majority of existing work in this area relies on classical non-deep techniques. Sun *et al*. [39] used images of different 2D face poses to create a 3D model. This was achieved by calculating the rotation and translation parameters with constrained independent component analysis and combining it with a prior 3D model for depth estimation of specific feature points. In a subsequent work [40], a nonlinear least-squares model was exploited to predict the depth of specific facial feature points, thereby inferring the 3D structure of the face. Both these methods used facial landmarks obtained by detectors for parameter initialization, making them highly dependent on landmark detection. Liu *et al*. [28] modelled image regions as superpixels and used optimization for depth estimation. In this context, the continuous variables encoded the depth of the superpixel while the discrete variables represented their internal relationships. In a later work, Zhu *et al*. [51] exploited the global structure of the scene by constructing a hierarchical representation of local, mid-level, and large-scale layouts. They modeled the problem as conditional Markov random field with variables for each layer in the hierarchy. In [22], Kong *et al*. mapped a 3D dataset to 2D images by sampling points from the dense 3D data and combining them with RGB channel information. They then exploited face Delaunay triangulation to create a structure of facial feature points. The similarity of the triangles among the test images and the training set allowed them to estimate depth.

A few methods have attempted to synthesize depth using deep learning architectures. Cui *et al*. [4] estimated depth from RGB using a multi-task approach consisting of face identification along with depth estimation. They also performed RGB-D recognition experiments to study the effectiveness of the estimated depth for the recognition task using an Inception-V2 [16] fusion network on the Lock3dFace and IIIT-D RGB-D datasets. Pini *et al*. [33] used a cGAN architecture for facial depth map estimation from monocular intensity images. Their method used co-registered intensity and depth images to train a generator in order to learn the relationship between RGB and depth images for face verification. Kwak *et al*. [24] proposed a solution based on

[1] https://github.com/hardik-uppal/teacher-student-gan.git

CycleGAN [50] for generating depth and image segmentation maps. To estimate depth information, the characteristics of input RGB images were maintained with the help of the consistency loss of CycleGAN. This was aided through a multi-task approach by generating segmentation maps for those RGB images which would further help the network to fill in depth information where it was ambiguous or hidden by overlapping of features of the image.

2.2. RGB-D Face Recognition

Early RGB-D facial recognition methods were proposed based on classical (non-deep) methods. Goswami *et al*. [9] fused visual saliency and entropy maps extracted from RGB and depth data. Histograms of oriented gradients were then used to extract features from image patches to then feed a classifier for identity recognition. Li *et al*. [26] used 3D point-cloud data to obtain a pose-corrected frontal view using a discriminant color space transformation. The corrected texture and depth maps were sparse approximated using separate dictionaries that were learned during the training phase. Hayat *et al*. [11] used a co-variance matrix representation on the Riemannian manifold to extract independent features from RGB and depth data, followed by an SVM classifier with score-level fusion to classify identities.

Recent methods have mainly focused on deep neural networks for RGB-D facial recognition. Chowdhury *et al*. [3] used Auto-Encoders to learn a mapping function between RGB and depth. The mapping function was then used to reconstruct depth images from the corresponding RGB to be used for identification. Zhang *et al*. [46] tackled the problem of multi-modal recognition using deep learning, focusing on joint learning of the CNN embedding to fuse the common and complementary information offered by the RGB and depth together effectively.

In [36], RGB, disparity maps, and depth images were independently used to fine-tune separate VGG-Face [32] models. The obtained embeddings were then fused to feed an SVM classifier for performing facial recognition. Jiang *et al*. [18] proposed an attribute-aware loss function for CNN-based facial recognition which aimed to regularize the distribution of learned representations with respect to soft-biometric attributes such as gender, ethnicity, and age, thus boosting recognition results. Lin *et al*. [27] proposed an RGB-D face identification method by introducing new loss functions, including associative and discriminative losses, which were then combined with softmax loss for training, showing boosted recognition results on the IIIT-D RGB-D dataset. Uppal *et al*. [42] proposed a two-level attention module to fuse RGB and depth modalities. The first attention layer selectively focused on the fused feature maps obtained by a convolutional feature extractor that were recurrently learned by an LSTM layer. The second attention layer then focused on the spatial features of those maps by applying attention weights using a convolution layer. In [42], the authors proposed an attention-based method in which the features of depth images allowed the network to focus on regions of the face in the RGB images that contained prominent person-specific information.

3. Method

3.1. Problem Formulation

We consider the problem of depth generation for a target dataset of RGB images $\{A_r\}_{i=1}^{M}$, whose distribution is $A_r \sim p_{target}(A_r)$, and have no corresponding depth information. Let's assume we are provided with an RGB-D dataset which we refer to as the teacher dataset $\{A_t, B_t\}_{i=1}^{N}$ with distribution $A_t, B_t \sim p_{train}(A_t, B_t)$, with A_t being an RGB image and B_t being the co-registered depth image. Our goal is to learn from the teacher dataset a mapping generator function G_{A2B} that can accurately generate an estimated depth image $\widetilde{B}_r$ for each target RGB image A_r.

3.2. Loss Formulation and Algorithm

Our end-to-end architecture TS-GAN consists of a teacher component and a student component. The aim of the teacher, which itself consists of a generator and a discriminator, is to learn a latent mapping between A_t and B_t. The student then refines the learned mapping for A_r by further training the generator, with the aid of another generator-discriminator pair. Figure 2 presents the TS-GAN architecture. For the teacher we create a mapping function, G_{A2B}:$A_t \rightarrow B_t$ along with a binary discriminator function $D_{depth}(.)$, which classifies whether the input is a real or fake (generated depth image). The loss $\mathcal{L}_{G_{A2B}}$ for the mapping function is then formulated as:

$$\mathcal{L}_{G_{A2B}} = \frac{1}{2}\mathbb{E}_{A_t \sim p_{train}(A_t)}[(D_{depth}(G_{A2B}(A_t)) - 1)^2], \quad (1)$$

where $\mathbb{E}_{A_t \sim p_{train}(A_t)}$ represents an RGB image sampled from $p_{train}(A_t)$, the distribution of RGB images in the teacher dataset.

The loss $\mathcal{L}_{D_{depth}}$ for the depth discriminator, whose goal is to differentiate between the ground truth and the hallucinated depth images, is:

$$\begin{aligned}\mathcal{L}_{D_{depth}} = &\frac{1}{2}\mathbb{E}_{B_t \sim p_{train}(B_t)}[(D_{depth}(B_t) - 1)^2] \\ &+\frac{1}{2}\mathbb{E}_{A_t \sim p_{train}(A_t)}[(D_{depth}(G_{A2B}(A_t))^2],\end{aligned} \quad (2)$$

where $\mathbb{E}_{B_t \sim p_{train}(B_t)}$ represents a depth image sampled from $p_{train}(B_t)$, the distribution of depth images in the teacher dataset.

The additional pixel loss, $\mathcal{L}_{pixel}$, between the hallucinated and ground truth depth can be formulated as:

$$\mathcal{L}_{pixel} = \frac{1}{n}\sum_{i=1}^{n} |(B_t)_i - G_{A2B}(A_t)_i| \,. \quad (3)$$

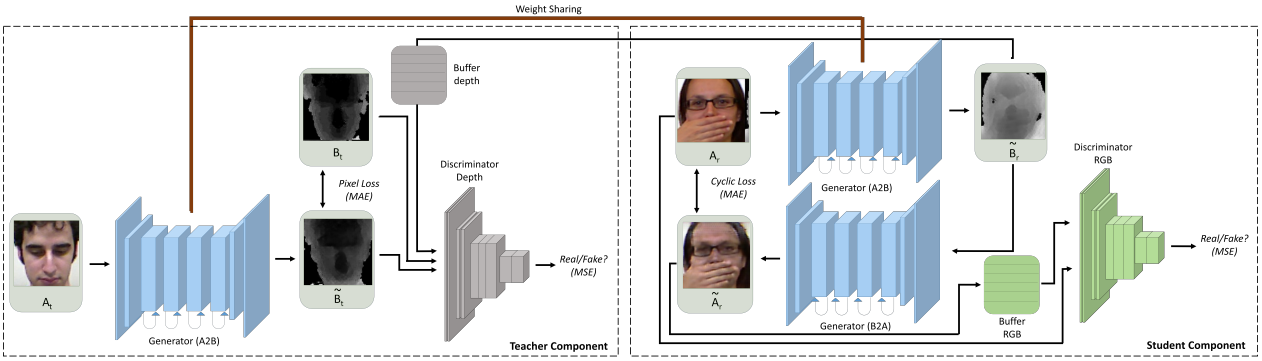

Figure 2. The architecture details for our proposed teacher-student adversarial network are presented. A_t and B_t refer to the co-registered RGB and depth images respectively, and $\widetilde{B_t}$ refers to the generated depth in the teacher component. A_r refers to the RGB image (when no corresponding depth is available), and and $\widetilde{A_r}$ denotes the reconstructed RGB image. $\widetilde{B_r}$ refers to the hallucinated depth generated by our model for the particular RGB image.

where n is the total number of pixels in an image.

The student component aims to convert a single RGB image A_r from the RGB dataset, for which no depth information is available, into a target depth image $\widetilde{B}_r$. This is done using the mapping function G_{A2B} from Eq. 1, along with an inverse mapping function G_{B2A}:$B_r \rightarrow A_r$, and a discriminator D_{RGB}. Loss $\mathcal{L}_{G_{B2A}}$ for the mapping function is formulated as:

$$\mathcal{L}_{G_{B2A}} = \frac{1}{2}\mathbb{E}_{A_r \sim p_{target}(A_r)} [(D_{RGB}(G_{B2A}(G_{A2B}(A_r))) - 1)^2], \tag{4}$$

where $\mathbb{E}_{A_r \sim p_{target}(A_r)}$ represents an RGB image sampled from $p_{target}(A_r)$, which is the distribution of RGB target dataset.

The loss $\mathcal{L}_{D_{RGB}}$ for the RGB discriminator whose goal is to discriminate between the ground truth RGB A_r and the generated RGB $\tilde{A}_r = G_{B2A}(G_{A2B}(A_r))$, is:

$$\mathcal{L}_{D_{RGB}} = \frac{1}{2}\mathbb{E}_{A_r \sim p_{target}(A_r)}[(D_{RGB}(A_r) - 1)^2] + \frac{1}{2}\mathbb{E}_{A_r \sim p_{target}(A_r)}[D_{RGB}(G_{B2A}(G_{A2B}(A_r)))^2]. \tag{5}$$

In addition to the supervisory signal from the discriminator, as discussed, we also employ another generator, G_{B2A}, to invert the mapping from the hallucinated depth back to RGB. This is done to preserve the identity of the subject and provide additional supervision in a cyclic-consistent way. Accordingly, we formulate the cyclic consistency loss as:

$$\mathcal{L}_{cyc} = \frac{1}{n}\sum_{i=1}^{n} |(A_r)_i - G_{A2B}(G_{B2A}(A_r))_i|, \tag{6}$$

The total loss for the teacher is then summarized as:

$$\mathcal{L}_{teach} = \mathcal{L}_{G_{A2B}} + \lambda_{pixel} \cdot \mathcal{L}_{pixel}, \tag{7}$$

where λ_{pixel} is the weighting parameter for the pixel loss, $\mathcal{L}_{pixel}$, described in Eq. 3.

Similarly, the total loss for the student component is summarized as:

$$\mathcal{L}_{student} = \mathcal{L}_{G_{A2B}} + \mathcal{L}_{G_{B2A}} + \lambda_{cyc} \cdot \mathcal{L}_{cyc}, \tag{8}$$

where λ_{cyc} is the weighting parameter for the cyclic loss, $\mathcal{L}_{cyc}$, described in Eq. 6.

The complete training process is listed in pseudocode in Algorithm 1. We first sample an RGB image A_t from $p_{train}(A_t)$ as input to the generator. The output of the generator is the estimated depth image $\widetilde{B_t}$, which is fed to the discriminator and classified as either real or fake. The discriminator is also trained with the corresponding ground truth depth image B_t, using the loss mentioned in Eq. 2. Apart from the adversarial loss, the training is facilitated with the help of pixel loss (Eq. 3), in the form of MAE loss, for which we define a weighting parameter λ_{pixel}.

After training the teacher, we sample an RGB image A_r from the target RGB data $p_{target}(A_r)$, and feed it as input to the generator that is shared between the student and the teacher. The estimated depth images $\widetilde{B_r}$ produced by this generator are then fed to the discriminator in the teacher network stream, thus providing a supervisory signal to generate realistic depth images. These hallucinated depth images are also fed to the inverse generator to transform the estimated depth back into estimated RGB $\widetilde{A_r}$ using the loss mentioned in Eq. 6. As discussed, this is done to preserve the identity information in the depth image while allowing for a more generalized mapping between RGB and depth to be learned through refinement of the original latent RGB-to-D mapping. An additional discriminator, which also follows a fully convolutional structure, is employed to provide an additional supervisory signal for the inverse generator to create realistic RGB images.

Algorithm 1: Teacher-student learning.

Input : teacher dataset $p_{train}(A_t, B_t)$, target RGB dataset $p_{target}(A_r)$, mapping generator function G_{A2B} and G_{B2A}, discriminators D_{RGB} and D_{Depth}, training configurations (loss weights: $\lambda_{pixel}, \lambda_{cyc}$; learning rates: $\alpha_{teach}, \alpha_{student}$; decay rate: β_{decay}; total epochs: N);

while *While n <N* **do**

- Sample $A_t, B_t \sim p_{train}(A_t, B_t)$;
- Compute loss $\mathcal{L}_{teach}(A_t, B_t; G_{A2B}, D_{Depth})$ using Eq. 7 and update G_{A2B};
- Compute loss $\mathcal{L}_{D_{depth}}(A_t, B_t; G_{A2B})$ using Eq. 2 and update D_{Depth};
- Sample $A_r \sim p_{target}(A_r)$;
- Compute loss $\mathcal{L}_{student}(A_r; G_{A2B}, G_{B2A}, D_{RGB})$ using Eq. 8 and update G_{A2B} and G_{B2A};
- Compute loss $\mathcal{L}_{D_{RGB}}(A_r; G_{A2B}, G_{B2A})$ using Eq. 5 and update D_{RGB};
- **if** $n >$epoch teacher **then** $\alpha_{teach} * \beta_{decay}$;
- **else** continue;
- **if** $n >$epoch student **then** $\alpha_{student} * \beta_{decay}$;
- **else** continue;

end

3.3. Implementation Details

Generator. We use a fully convolutional structure for the generator inspired by [50], where an input image of size $128 \times 128 \times 3$ is used to output a depth image with the same spatial dimensions. The encoder part of the generator contains three convolution layers with ReLU activation, where the number of feature maps is gradually increased $(64, 128, 256)$ with a kernel size of 7×7 and a stride of 1 for the first layer. Subsequent layers use a kernel size of 3×3 and a stride of 2. This is followed by 6 residual blocks, consisting of 2 convolution layers each with a kernel size of 3×3, a stride of 2, and 256 feature maps. The final decoder part of the generator follows a similar structure, with the exception of using de-convolution layers for upsampling instead of convolution, with decreasing feature maps $(128, 64, 3)$. The last de-convolution layer which is used to map the features back to images uses a kernel size of 7×7 and a stride of 1, the same as the first layer of the encoder, but with a tanh activation.

Discriminator. We use a fully convolutional architecture for the discriminator, with an input of size $128 \times 128 \times 3$. The network uses 4 convolution layers, where the number of filters gradually increase $(64, 128, 256, 256)$, with a fixed kernel of 4×4 and a stride of 2. All the convolution layers use Instance normalization and leaky ReLU activations with a slope of 0.2. The final convolution layer uses the same parameters, but with only 1 feature map.

Training. For stabilizing the model, we use the strategy proposed in [37], updating the discriminators using images from a buffer pool of 50 generated images rather than the ones immediately produced by the generators. Our proposed network is trained from scratch on an Nvidia RTX 2080Ti GPU, using TensorFlow 2.2. We use Adam optimizer and a batch size of 1 as done in [50]. Additionally, we use two different learning rates of 0.0002 and 0.000002 for the teacher and student components respectively. Following the suggestions in [45], we start decaying the learning rate for the teacher on the 25^{th} epoch with a decay rate 0.5, sooner than the student, where the learning rate decay starts after the 50^{th} epoch. The weights λ_{cyc} and λ_{pixel} are empirically determined to be 5 and 10, respectively.

4. Experiments

4.1. Datasets

CurtinFaces [25] is a common RGB-D face dataset which contains over 5000 co-registered RGB and depth image pairs from 52 subjects, captured with a Microsoft Kinect [49]. It has been recorded with varying poses, expressions, and under multiple illumination variations.

IIIT-D RGB-D [9, 10] contains 4605 RGB-D images from 106 subjects captured using a Microsoft Kinect in two acquisition sessions. Each subject has been captured under normal illumination conditions with variations in pose, expression, and eyeglasses. Each image in the dataset is pre-cropped around the face.

EURECOM KinectFaceDb [30] contains RGB-D face images from 52 people (14 female and 38 male) obtained by a Microsoft Kinect. The data has been captured in 2 different sessions with variations in expression, pose, illumination, and occlusion (a total of 18 images per subject).

Labeled Faces in-the-wild (LFW) [15] contains more than 13,000 face images collected from the Internet. Each face has been labeled with the name of the person, with 62 subjects having more than 20 images.

4.2. Evaluation

Protocols. In the training phase, we use the CurtinFaces dataset to train the teacher in order to learn a strict latent mapping between RGB and depth. We choose this dataset as it contains minimal noise among the RGB-D datasets considered in this study, and contains over 5000 co-registered RGB-D images making it the largest. We use its RGB and ground-truth depth images as A_t and B_t respectively (see Section 3.2). To train the student, we use the training subsets of the RGB images from IIIT-D RGB-D and EURECOM KinectFaceDb. IIIT-D RGB-D has a predefined protocol with a 5-fold cross-validation strategy, to which we strictly adhere. For EURECOM KinectFaceDb, we divide the data into a 50-50 split between the training and testing sets, resulting in a total of 468 images in each set. In the case of the in-the-wild LFW RGB dataset, we

utilize 11,953 images for training the generator, and keep the rest of the images for recognition experiments.

For the testing phase of our experiments, we use the trained generator from the student to generate the hallucinated depth images for each RGB image in the test sets. We then further use the RGB and depth images to train the various recognition networks mentioned in Section 5.2. For RGB-D datasets, we train the recognition networks on the training sets using the RGB and hallucinated depth images, and evaluate the performance on the test sets. Concerning the LFW dataset, in the testing phase, we use the remaining 20 images from each of the 62 identities that are not used for training. We then use the output RGB and hallucinated depth images as inputs for the recognition experiment.

Metrics. We first verify the quality of our depth generation against other generators using pixel-wise quality assessment metrics with respect to the original co-registered ground truth depths. These metrics includes pixel-wise absolute difference, L1 norm, L2 norm, and Root Mean Squared Error (RMSE) [5, 33]. We also use a threshold metric (δ) [5], defined as $\%$ of y_i s.t. $\max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) = \delta < val.$, which measures the percentage of pixels under a certain error threshold, thus providing a similarity score. In this metric, y_i and y_i^* represent pixel values in ground truth and hallucinated depths respectively, and val denotes the threshold error value which has been set to 1.25 as suggested in [5]. We also use the Fréchet Inception Distance (FID) score [13] as a measure of similarity between ground truth depth and synthetic depth images.

Face Recognition. The aim of this study is to use the hallucinated modality to boost recognition performance. As we aim to present results with no dependency on a specific recognition architecture, we use a diverse set of standard deep networks, notably VGG-16 [38], inception-v2 [16], ResNet-50 [12], and SE-ResNet-50 [14] in our evaluation. We report the rank-1 identification results with and without ground truth depth for RGB-D datasets as well as the results obtained by the combination of RGB and our hallucinated depth images. For LFW RGB dataset, we naturally do not have ground truth depths, so we only present the identification results with and without our hallucinated depth. We also use different strategies, including feature-level fusion, score-level fusion, two-level attention fusion [41], and depth-guided attention [42], when combining RGB and depth images.

5. Performance

5.1. Quality Assessment

We first compare the performance of TS-GAN with alternative depth generators, namely Fully Convolutional Network (FCN) [4], image-to-image translation cGAN [33], and CycleGAN [24] To this end, we perform experiments

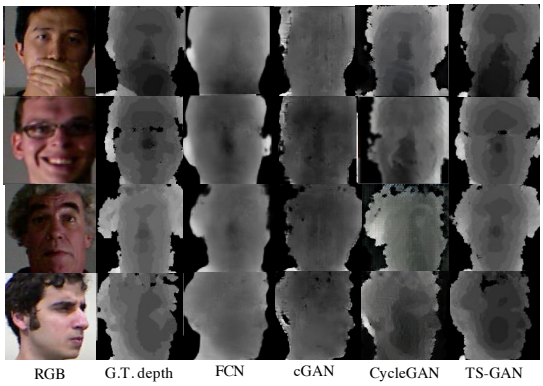

Figure 3. Several input RGB test samples from the CurtinFaces dataset along with ground truth (G.T.) co-registered depth images, and synthesized depth images generated by various state-of-the-art alternatives and our proposed method are presented.

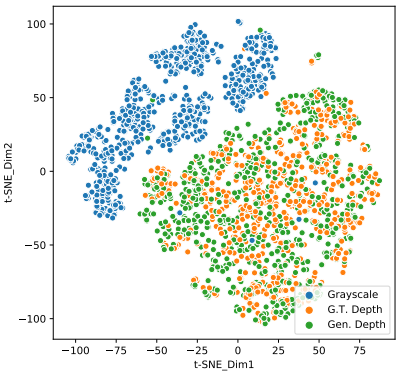

Figure 4. A t-SNE visualization of input RGB images (converted to grayscale), Ground Truth (G.T.) depth, and the output depth Hallucinated (Hal.) by TS-GAN.

on the CurtinFaces dataset, where we use 47 out of the 52 subjects for training the generator, and use the remaining 5 subjects for generating depth images to be used for quality assessment experiments. Figure 3 shows depth generated by the alternative methods as well as our TS-GAN on some of the test subjects. As can be seen, our method is able to generate realistic depth images which appear very similar to the ground truth depth images.

In Figure 4 we present a t-SNE [43] visualization of embeddings generated by a ResNet-50 network for a number of RGB samples from CurtinFaces (converted to grayscale in order for color to not be considered a factor), ground truth depth images, and hallucinated depth images by TS-GAN. This figure demonstrates a very high degree of overlap between the ground truth and generated depth images, thus depicting their similarity.

Table 1 shows the results for pixel-wise objective metrics (Section 4.2). For the first four metrics namely absolute

Table 1. Comparisons of image quality metrics between our method and other depth generation methods.

Metrics	FCN [4]	cGAN [33]	CycleGAN [24]	Ours (TS-GAN)
Abs. Diff. ↓	**0.0712**	0.0903	0.1037	0.0754
L1 Norm ↓	0.2248	0.2201	0.2387	**0.2050**
L2 Norm ↓	89.12	89.05	90.32	**82.54**
RMSE ↓	0.3475	0.3474	0.3542	**0.3234**
$\delta(1.25)$ ↑	64.31	64.27	65.76	**69.02**
$\delta(1.25^2)$ ↑	81.66	82.08	82.56	**87.20**
$\delta(1.25^3)$ ↑	94.33	95.10	95.63	**97.54**
FID ↓	17.72	16.39	16.13	**14.67**

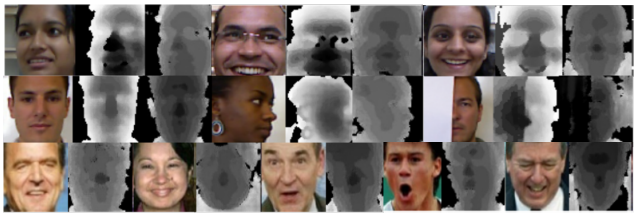

Figure 5. The first two rows show samples of the RGB-D datasets (IIIT-D and EURECOM KinectFaceDb). The first column shows RGB images, the second column shows the ground truth depth, and the third column shows the hallucinated depth. In the third row LFW samples are presented where the first column shows the RGB images while the second column shows the hallucinated depth.

difference, L1 Norm, L2 Norm, and RMSE, lower values indicate better image quality. It can be observed that our proposed method mostly outperforms the other methods, the single exception being the absolute difference metric in which FCN shows slightly better performance. A potential reason for this anomaly is that FCN only uses one loss function that aims to minimize the absolute error between the ground truth and the generated depth, naturally resulting in minimal absolute difference error. For the threshold metric δ, the higher percentage of pixels under the threshold error value of 1.25 achieved by our method represents better spatial accuracy for the generated depth images. Lastly, the lower obtained FID scores indicate that the proposed method images are most similar to the ground truth depth samples.

In order to show the generalization of our generator when applied to the target datasets (mentioned in Section 3) for testing, hallucinated depth samples for IIIT-D and EURECOM RGB-D datasets are shown in Figure 5 (top and middle rows). The first and second columns show the input RGB images and the ground truth depth image corresponding to the RGB image, while the third column shows the generated depth images. As can be seen, our methods can adopt to different poses, expressions and occlusions present in the target datasets. The bottom row in this figure shows the depth generated for the in-the-wild LFW RGB dataset, where our method is able to adopt to the non-frontal and unnatural poses which are not present in the constrained, lab-acquired RGB-D datasets.

Table 2. IIIT-D and EURECOM KinectFaceDb rank-1 recognition results. $\widetilde{\text{D}}$ denotes the hallucinated depth using TS-GAN.

Dataset	Model	Accuracy				
		RGB	RGB + D		RGB + $\widetilde{\text{D}}$	
		–	Feat. Fusion	Score Fusion	Feat. Fusion	Score Fusion
IIIT-D	VGG-16	94.1%	**95.4%**	94.4%	**95.4%**	94.1%
	Inception-v2	95.0%	**96.5%**	95.0%	96.1%	95.9%
	ResNet-50	95.8%	96.9%	95.9%	**97.1%**	96.1%
	SE-ResNet-50	97.8%	**98.9%**	97.9%	98.6%	97.6%
	Two-level att. [41]	–	**99.4%**	–	99.1%	–
	Depth-guid. att. [42]	–	**99.7%**	–	**99.7%**	–
EURECOM	VGG-16	83.6%	**88.4%**	84.5%	88.3%	84.2%
	Inception-v2	87.5%	**90.3%**	86.9%	90.1%	87.9%
	ResNet-50	90.8%	92.1%	91.0%	**92.2%**	90.7%
	SE-ResNet-50	91.3%	93.1%	91.6%	**93.2%**	91.5%
	Two-level att. [41]	–	92.0%	–	**92.3%**	–
	Depth-guid. att. [42]	–	**93.1%**	–	92.7%	–

Table 3. EURECOM KinectFaceDb pose and occlusion test set recognition.

Test set	Model	Accuracy		
		RGB	RGB + D	RGB + $\widetilde{\text{D}}$
Left Pose Set	VGG-16	75.2%	**77.4%**	77.2%
	Inception-v2	75.8%	**78.1%**	77.6%
	ResNet-50	77.4%	80.4%	**80.5%**
	SE-ResNet-50	79.2%	80.8%	**81.1%**
	Two-level att.	–	**81.6%**	81.3%
	Depth-guid. att.	–	82.5%	**82.7%**
Right Pose Set	VGG-16	74.8%	**77.6%**	77.5%
	Inception-v2	75.9%	**78.6%**	78.4%
	ResNet-50	77.2%	80.1%	**80.3%**
	SE-ResNet-50	78.9%	80.4%	**80.7%**
	Two-level att.	–	**81.9%**	81.5%
	Depth-guid. att.	–	**82.6%**	82.3%
Occlusion Set	VGG-16	84.8%	**87.4%**	87.2%
	Inception-v2	86.2%	88.3%	**89.8%**
	ResNet-50	88.9%	90.1%	**90.8%**
	SE-ResNet-50	90.8%	92.2%	**92.5%**
	Two-level att.	–	**92.5%**	**92.5%**
	Depth-guid. att.	–	**93.8%**	93.2%

5.2. Recognition Results

RGB-D Datasets. To demonstrate the effectiveness of the hallucinated depth for face recognition, the mapping function (Eq. 1) is used to estimate the corresponding depth images for the RGB images, both of which are used as input to the recognition pipeline. Table 2 shows the rank-1 recognition results on the IIIT-D and KinectFaceDb datasets using the four networks discussed earlier. We have considered different fusion strategies as well as two recent attention-based RGB-D solutions [41, 42] as mentioned in Section 4.2. It can be observed that the fusion of RGB and the depth hallucinated using TS-GAN constantly provides better results across all the CNN architectures, when compared to using only the RGB images.

For further comparison, we also perform recognition with RGB and the ground truth depth using the same pipelines. For the IIIT-D dataset, recognition with RGB and generated depth leads to comparable results to that with RGB and ground truth depth images. Concerning the EURECOM KinectFaceDb dataset, the results also show that our generated depth provide added value to the recognition

Table 4. LFW rank-1 recognition results. $\widetilde{\mathrm{D}}$ denotes the hallucinated depth using TS-GAN.

Model	Accuracy		
	RGB	RGB + $\widetilde{\mathrm{D}}$	
	–	Feature Fusion	Score Fusion
VGG-16	75.3%	**78.7%**	76.1%
Inception-v2	78.1%	**80.5%**	78.4%
ResNet-50	81.8%	**84.1%**	81.7%
SE-ResNet-50	83.2%	**85.6%**	83.2%
Two-level att. [41]	–	**84.7%**	–
Depth-guided att. [42]	–	**85.9%**	–

Table 5. Ablation study on IIIT-D and EURECOM KinectFaceDb.

Ablation Model	Classifier	IIIT-D Accuracy	KinectFaceDb Accuracy
Teacher	VGG-16	**95.4%**	85.7%
	Inception-v2	95.0%	88.6%
	ResNet-50	96.6%	91.3%
	SE-ResNet-50	98.4%	91.9%
Teacher's A2B Gen.	VGG-16	95.1%	87.8%
	Inception-v2	96.0%	88.2%
	ResNet-50	96.7%	90.6%
	SE-ResNet-50	98.5%	92.2%
Teacher-Student (TS-GAN)	VGG-16	**95.4%**	**88.3%**
	Inception-v2	**96.1%**	**90.1%**
	ResNet-50	**97.1%**	**92.2%**
	SE-ResNet-50	**98.6%**	**93.2%**

pipeline as competitive results (slightly below) to that of RGB and ground truth depth are achieved. Interestingly, in some cases for both IIIT-D and KinectFaceDb, our hallucinated depth images even provide superior performance over the ground-truth depth images. This is most likely due to the fact that some depth images available in the IIIT-D and KinectFaceDb datasets are noisy, while our generator can provide cleaner synthetic depth images as it has been trained on higher quality depth images available in the CurtinFaces dataset. Finally, to test the robustness of our approach on variations in pose and occlusions, we perform experiments using the EURECOM KinectFaceDb dataset. The results presented in Table 3 indicate that our TS-GAN results in high quality depth images even with variations in pose and occlusions, as evidenced by the high recognition rates.

RGB Dataset. Table 4 presents the recognition results on the in-the-wild LFW dataset, where the results are presented both with and without our hallucinated depth images. We observe that the hallucinated depth significantly improves the recognition accuracy across all the CNN architectures, with 3.4%, 2.4%, 2.3%, 2.4% improvements for VGG-16, Inception-v2, ResNet-50, and SE-ResNet-50 respectively. The improvements are more obvious when considering the state-of-the-art attention-based methods, clearly demonstrating the benefits of our synthetic depth images to improve recognition accuracy.

5.3. Ablation study

To evaluate the impact of each of the main components of our solution, we perform ablation experiments by systematically removing them. First, we remove the student component, leaving just the teacher. Next, we remove the discriminator from the teacher leaving only the A2B generator as discussed in Section 3 (also see Figure 2). The results are presented in Table 5 and compared to our complete TS-GAN solution. The presented recognition results are obtained using a feature-level fusion scheme to combine RGB and hallucinated depth images. The results show that performance suffers by the removal of each component for all four CNN architectures, demonstrating the effectiveness of our approach.

Figure 6. A few samples of failed depth hallucination.

6. Limitations

Although our proposed method provides impressive results, high quality depth images can not be generated in some cases. Our proposed method has been successful in adopting to various non-frontal poses and expressions as can be seen in Figure 5, however, it maintains some sensitivity to the diversity of the training data. For instance, gray-scale images were not used to train the generator, and hence adopting to them proves difficult for the generator as seen in Figure 6. Our method also fails to generate high quality depth images for very low-resolution images and multiple faces in the same image. To mitigate these problems, we could create a larger and more diverse training set to include these variations during training which could help the generator with better generalization.

7. Conclusion

In this paper, we propose a novel teacher-student adversarial architecture for depth generation from RGB images, called TS-GAN, to boost the performance of facial recognition systems. The teacher component of our method consisting of a generator and a discriminator learns a strict latent mapping between RGB and depth image pairs following a supervised approach. The student, which itself consists of a generator-discriminator pair along with the generator shared with the teacher, then refines this mapping by learning a more generalized relationship between the RGB and depth domains for samples without corresponding co-registered depth images. Comprehensive experiments on three public face datasets show that our method outperformed other depth generation methods.

Acknowledgements. The authors would like to thank Irdeto Canada Corporation and the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding this research.

References

[1] Yael Adini, Yael Moses, and Shimon Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.

[2] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3d: Stereo depth estimation via binary classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1600–1608, 2020.

[3] Anurag Chowdhury, Soumyadeep Ghosh, Richa Singh, and Mayank Vatsa. RGB-D face recognition via learning-based reconstruction. In *International Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2016.

[4] Jiyun Cui, Hao Zhang, Hu Han, Shiguang Shan, and Xilin Chen. Improving 2D face recognition via discriminative face depth estimation. In *International Conference on Biometrics*, pages 140–147, 2018.

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 27:2366–2374, 2014.

[6] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[7] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

[9] Gaurav Goswami, Samarth Bharadwaj, Mayank Vatsa, and Richa Singh. On RGB-D face recognition using Kinect. In *International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6. IEEE, 2013.

[10] Gaurav Goswami, Mayank Vatsa, and Richa Singh. RGB-D face recognition with texture and attribute features. *IEEE Transactions on Information Forensics and Security*, 9(10):1629–1640, 2014.

[11] Munawar Hayat, Mohammed Bennamoun, and Amar A El-Sallam. An RGB–D based image set classification for robust face recognition from kinect data. *Neurocomputing*, 171:889–900, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[15] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[18] Luo Jiang, Juyong Zhang, and Bailin Deng. Robust RGB-D face recognition using attribute-aware loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2552–2566, 2020.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[22] Dezhi Kong, Yang Yang, Yun-Xia Liu, Min Li, and Hongying Jia. Effective 3d face depth estimation from a single 2d face image. In *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*, pages 221–230. IEEE, 2016.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[24] Dong-hoon Kwak and Seung-ho Lee. A novel method for estimating monocular depth using cycle gan and segmentation. *Sensors*, 20(9):2567, 2020.

[25] Billy Li, Ajmal Mian, Wanquan Liu, and Aneesh Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *IEEE Workshop on Applications of Computer Vision*, 2013.

[26] Billy YL Li, Ajmal S Mian, Wanquan Liu, and Aneesh Krishna. Face recognition based on Kinect. *Pattern Analysis and Applications*, 19(4):977–987, 2016.

[27] Tzu-Ying Lin, Ching-Te Chiu, and Ching-Tung Tang. RGB-D based multi-modal deep learning for face identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1668–1672, 2020.

[28] Miaomiao Liu, Mathieu Salzmann, and Xuming He. Discrete-continuous depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723, 2014.

[29] Mostafa Mehdipour Ghazi and Hazim Kemal Ekenel. A comprehensive analysis of deep learning based representation for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 34–41, 2016.

[30] Rui Min, Neslihan Kose, and Jean-Luc Dugelay. Kinectfacedb: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(11):1534–1548, 2014.

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[32] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[33] Stefano Pini, Filippo Grazioli, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. Learning to generate facial depth maps. In *2018 International Conference on 3D Vision (3DV)*, pages 634–642. IEEE, 2018.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[35] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1–9. IEEE, 2016.

[36] A. Sepas-Moghaddam, P. L. Correia, K. Nasrollahi, T. B. Moeslund, and F. Pereira. Light field based face recognition via a fused deep representation. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2018.

[37] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2107–2116, 2017.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] Zhan-Li Sun and Kin-Man Lam. Depth estimation of face images based on the constrained ica model. *IEEE Transactions on Information Forensics and Security*, 6(2):360–370, 2011.

[40] Zhan-Li Sun, Kin-Man Lam, and Qing-Wei Gao. Depth estimation of face images using the nonlinear least-squares model. *IEEE Transactions on Image Processing*, 22(1):17–30, 2012.

[41] Hardik Uppal, Alireza Sepas-Moghaddam, Michael Greenspan, and Ali Etemad. Two-level attention-based fusion learning for rgb-d face recognition. *International Conference of Pattern Recognition*, 2020.

[42] Hardik Uppal, Alireza Sepas-Moghaddam, Michael Greenspan, and Ali Etemad. Depth as attention for face representation learning. *IEEE Transactions on Information Forensics and Security*, 16:2461–2476, 2021.

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[44] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *International conference on 3d vision*, pages 248–257. IEEE, 2018.

[45] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan Yuille. Knowledge distillation in generations: More tolerant teachers educate better students. *arXiv preprint arXiv:1805.05551*, 2018.

[46] Hao Zhang, Hu Han, Jiyun Cui, Shiguang Shan, and Xilin Chen. RGB-D face recognition via deep complementary and common feature learning. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 8–15, 2018.

[47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.

[48] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE Pattern Analysis and Machine Intelligence on Pattern Analysis and Machine Intelligence*, 41(3):611–624, 2018.

[49] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 19(2):4–10, 2012.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[51] Wei Zhuo, Mathieu Salzmann, Xuming He, and Miaomiao Liu. Indoor scene structure analysis for single image depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 614–622, 2015.