Supplementary Material

In the supplementary material, we give additional information for our method. In Section 1 we provide more details on the thermalization including implementation details and an extended ablation study. In Section 2, we add details on landmarker and label adaptation implementation. In Sections 3-5 we discuss limitations and provide additional result images of the datasets.

1. Thermalization

1.1. Reference Temperature Values

As described in the main document, we train two versions of the thermalizer T_θ to model facial temperature variations under different environmental conditions. For that purpose, we use two sets of reference temperatures, a 'cold' and a 'warm' condition, for the different facial regions that guide the segmentation-based regularizer. These are presented in Table 1. Thermal facial contrast is increased and overall body temperature is decreased for the 'cold' condition in comparison to the 'warm' condition which is in line with empirical findings [1].

Segmentation	'Cold'	'Warm'
Background, Glasses	<20	<20
Skin	33	35
Nose	31.5	35
Eyes	34	35
Brows	31	34
Ears	32	35
Mouth Interior	35	35
Lips	32.5	35
Neck	34	35
Hair	30	30
Beard	31	32
Clothing	30	32
Headwear, Facewear	28	28

Table 1. Reference temperatures in Celsius for our segmentation-based regularizer for the 'cold' and the 'warm' setup. Note that our pixel range goes from 20 °C to 40 °C.

1.2. Thermalization Implementation Details

For the patch-based regularizers, we use random batches with patch size of 8. We disregard the background by excluding synthetic patches based on the ground truth background segmentation and completely black real patches. For our multi-scale approach, we sum the regularizer over 5 scales with a downsampling factor of 0.5. Again, facial temperatures depend on the surrounding temperature. Thus, we train a 'warm' and a 'cold' model with different reference temperature values for the segmentation-based regularizers. For data augmentation, we use the same random rotations and cropping for the natural RGB and thermal images. Moreover, we apply random color changes, blurring, and shadow augmentations [14] exclusively to the natural RGB images. Next, we apply random rotations and cropping for the synthetic RGB images. Here, we also fill holes in the original 'glasses' segmentation masks showing outlines of the frame only to highlight transparent, but heat-blocking glass or plastic. Lastly, we replace the original background with a black background based on the known semantic segmentation. We use a U-Net T_θ with a Resnet34 encoder pre-trained on ImageNet [3] and train it for 10 FAKE epochs with SEJONG batches of size 64 and FAKE batches of size 64. This corresponds to approximately 100 SEJONG epochs. Further, we use an Adam optimizer with an initial learning rate of 0.001 which we reduce to 0.0001 after 4 epochs. Based on a random split, we use 80% of the SEJONG data for training and all available FAKE data. Also, we use the *geomloss* [4] implementation for the Wasserstein patch loss with $\lambda_E = 0.1^6$. We normalize the squared error loss $\|\cdot - \cdot\|_2^2$ by dividing with the image dimension, here 256^2. Based on the 5 scales and patch dimension 8^2, we set $\lambda_W = 0.01C$ for our final model with the normalization constant $C = (5 \cdot 8^2)^{-1}$. We set $\lambda_R = 1$. Note that the choice of $\lambda_R = 1$ and C are motivated to control the value range. The (normalized) MSE data fidelity term $\|\cdot - \cdot\|_2^2$ which we evaluate on normalized SEJONG images takes values in the range $[0, 1]$ for arbitrary images with pixel range $[0, 1]$ due to dimensional normalization. To have a similar value range for the evaluated FAKE images, the patch-based regularizer W takes values in $[0, 1]$ for $\lambda_W = C$ with arbitrary normalized images. Moreover, the segmentation-based regularizer R also takes values in $[0, 1]$ for such images given arbitrary normalized reference temperature values in $[0, 1]$ for $\lambda_R = 1$.

1.3. SEJONG Dataset

The SEJONG dataset illustrates the impact of various disguises, including glasses, wigs, and fake beards. Each subject in the dataset is presented with different disguises. As a result, it includes a lot of clothing and hairstyle variations. Most participants have a Southeast Asian or Central Asian ethnic background, but people of other ethnicities are included too. The number of participants identifying as male or female is balanced. This makes it an attractive dataset candidate for thermalization training that is supposed to generalize to a large variety of subjects. Due to reasons of data privacy, we refer to the original publication [2] and have to abstain from showing additional SEJONG images.

1.4. Generalization to Out-of-Lab Conditions

The success of neural networks in the last two decades has been tremendous, especially in the imaging domain. Nevertheless, a common empirical finding is the limited capability of neural networks to generalize to new settings [16]. Often, this limitation is caused by biased datasets. Particularly in the biomedical domain, data acquisition is a valuable task. However, due to real-world restrictions, data is often acquired in laboratory conditions. Acquiring paired RGB and thermal facial images requires a calibrated multimodal camera setup. Most multimodal facial datasets are restricted to frontal views with relatively neural expressions and frontal lighting, e.g. [2, 5, 17]. As a direct consequence, most Thermal2RGB research has focused on learning the transformation purely for *frontal images with frontal lighting at room temperature* [15, 20, 23]. To our knowledge, our regularized model is the first facial model to promote the explicit generalization of a learned thermal transformation to new poses, facial expressions, and lighting conditions and to simulate distinct temperature conditions. However, due to a lack of paired multimodal 'in-the-wild' facial datasets, we have to partially resort to metrics for unsupervised image translation, i.e., the FID [8]. To additionally visualize this result, we show results for the TUFTS [17] dataset containing RGB and thermal images. Again, we find that paired RGB and images are only available for frontal views. We present the results of applying our baseline model without regularization ($\lambda_W = 0$, $\lambda_R = 0$), a Pix2Pix model, and our final 'cold' model to the dataset subset with paired RGB images in Figure 1. We chose the 'cold' model because the images were recorded at room temperature. We see only a marginal impact of our regularization on the prediction. However, the TUFTS dataset additionally contains RGB images without paired thermal images recorded from different angles. Therefore, we also apply both models to random images taken from a fixed side angle. The result is displayed in Figure 2. We see that the unregularized model generates various large facial artifacts whereas our final model contains almost no facial artifacts. This shows

Figure 1. Thermalization results for *frontal* RGB images *with* paired thermal images from the TUFTS [17] database without regularization ($\lambda_W = 0$, $\lambda_R = 0$), Pix2Pix and our final model ($\lambda_W = 0.001C$, $\lambda_R = 1$) (top to bottom).

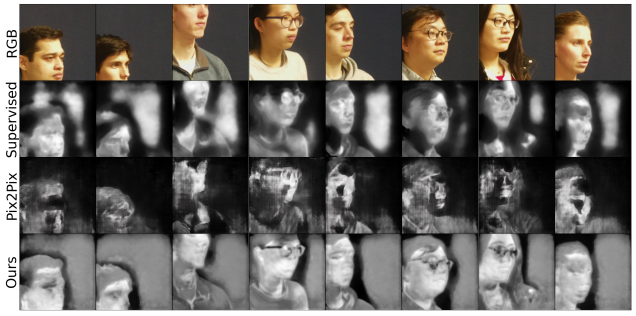

Figure 2. Thermalization results for *side* RGB images *without* paired thermal images from the TUFTS [17] database without regularization ($\lambda_W = 0$, $\lambda_R = 0$), Pix2Pix and our final model ($\lambda_W = 0.001C$, $\lambda_R = 1$) (top to bottom).

that the main limitation of training RGB2Thermal and Thermal2RGB models is the lack of paired 'in-the-wild' multimodal images. Our regularization allows us to overcome this limitation.

1.5. Thermalization Comparison Details

We used the official PyTorch implementation for all compared models. We trained Pix2Pix and all SEJONG images and all unsupervised images using all ∼10k SEJONG images and 10k FAKE (RGB) images. We used default hyperparameters and trained all models for 10 epochs with a constant learning rate and an additional 10 epochs with a linearly decreasing learning rate. All models were trained and evaluated with a resolution of 256 for the FID and the MSE. However, our models were evaluated on a resolution of 512 and the output was downscaled to 256 to be in line with the final T-FAKE dataset. We display examples in Figure 3. For the MSE comparison, we removed the background because we use random background augmentations for our final landmarker training. Here, we removed the background

Figure 3. Comparison of FAKE images thermalized with Pix2Pix [10], CycleGAN [24], CUT [18], QS-Attn [9], our supervised baseline and our final model (top to bottom, background removed).

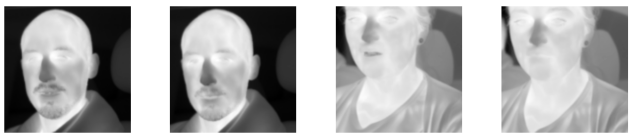

Figure 4. Thermal ground truth samples from the DRIVE-IN [5] dataset with side profile.

by masking all predictions based on a 21 °C threshold based on the ground truth. For the FID evaluation, we randomly choose the 'warm' or the 'cold' variant for each image for our model. For the MSE, we average the results for 'warm' and 'cold' variants. Due to data privacy reasons, we are only able to present ground truth thermal images for two persons from the DRIVE-IN dataset [5], see Figure 4.

1.6. Extended Thermalization Ablation Study

Given that the regularizers are solely defined for the synthetic images, we fix $\lambda_T = 1$ to ensure that the regularizer is on the same scale as the MSE of the real images and vary λ_W for our ablation. We train all models with the same setup and random seeds. The FID implementation is the default *PyTorch-Ignite* [6] implementation. Here, we extend the ablation study table in the main document which only displays the best result for $\lambda_W = 0.01C$. In particular, we display the results of a grid search for λ_W and λ_R in Table 2. As described in the main document, we calculate the mean FID and its standard deviation for five different subsets of the T-FAKE dataset. Moreover, we use the same

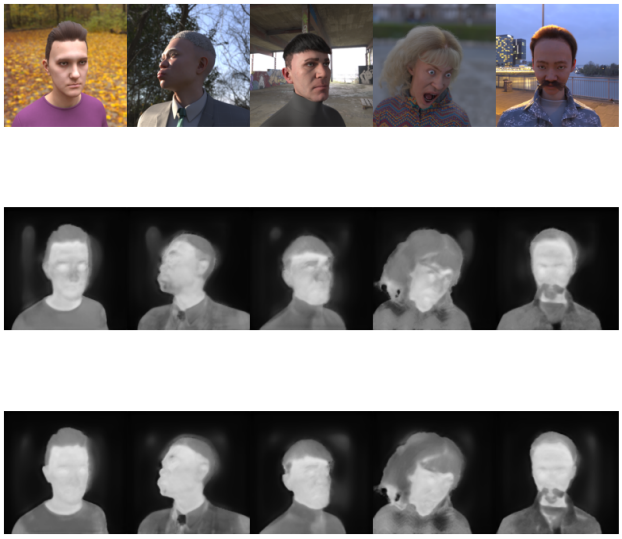

Figure 5. T-Fake samples with original images (first row), 'cold' images' (second row), and 'warm' images' (third row). Note the diminished contrasts of the noses and the checks.

Regularization	$\lambda_R = 1$	$\lambda_R = 0$
$\lambda_W = 1C$	.1665 ± .0030	.3375 ± .0056
$\lambda_W = 0.1C$	.1753 ± .0018	.3654 ± .0032
$\lambda_W = 0.01C$	.1598 ± .0041	.3146 ± .0092
$\lambda_W = 0$	.1706 ± .0029	.5028 ± .0054

Table 2. Impact of the regularization parameters on the perceptual quality measured with the FID (↓) with $C = (5 \cdot 8^2)^{-1}$.

setup to compare the perceptual quality of the 'cold' and the 'warm' setup of our final T-FAKE dataset, see Table 3. Figure 5 shows T-FAKE samples with the 'cold' and 'warm' variants. In addition, we present some samples generated with different regularization configurations in Figure 6 for the 'cold' setup and in Figure 7 for the 'warm' setup. Here, we can visually see the impact of the different regularizers.

According to the FID, both regularizers have a positive impact. The segmentation-based regularizer greatly boosts the perceptual quality, while the effect of the patch-based regularizer is smaller. The optimal FID value is obtained for $\lambda_W = 0.01C$ and $\lambda_T = 1$, the parameters of our final model. The perceptual quality of the figures seems in line with the FID. A closer look at the last two rows in both figures shows that the segmentation-based regularizer alone can lead to smoothed facial areas and overly exaggerated differences on the edges of the semantic segmentation. This becomes more apparent for the 'cold' setup as it leads to more thermal contrast within the face, see Figure 6. The FID shows only a small difference between the 'cold' and 'warm' variants. The 'warm' variant displays slightly lower FID values. For a visual comparison of the T-FAKE images

Setup	'Cold'	'Warm'
FID ↓	.1577 ± .0024	.1685 ± .0111

Table 3. Perceptual comparison of thermal setups 'cold' and 'warm' using the FID.

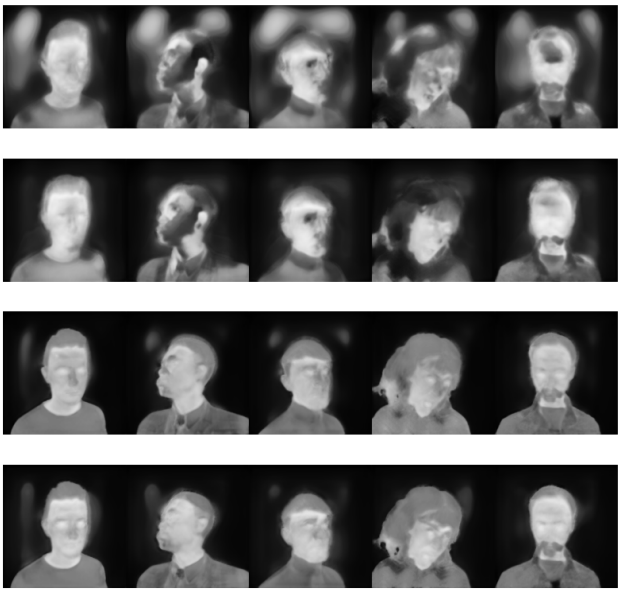

Figure 6. Regularization impact on images for 'cold' setup: No regularization ($\lambda_W = 0$, $\lambda_R = 0$), only patch-based ($\lambda_W = 0.01C$, $\lambda_R = 0$), only segmentation-based ($\lambda_W = 0$, $\lambda_R = 1$), and final model ($\lambda_W = 0.01C$, $\lambda_R = 1$) (top to bottom).

with real thermal images, we refer to Fig. 8.

2. Landmarking

2.1. CHARLOTTE Dataset

The CHARLOTTE dataset contains thermal images with varying thermal conditions, various head positions, and multiple camera distances. Moreover, it contains information about the thermal sensation of the subjects. We refer to Fig. 8 for a visualization of some thermal CHARLOTTE images without landmarks.

2.2. Landmarking Implementation Details

For training, we include random landmark positions on a texture dataset [7] as negative examples into the dataset to increase the learned uncertainty σ^2 on images without faces. On thermal images, we fill in the background with random textures from the texture dataset [7] with a probability of 0.25. During inference, we use a multi-scale sliding window evaluation to generalize our model to varying image sizes and face scales. We downsample iteratively with a factor of 0.75 until the height or the width reaches 224. For each image scale, we run our model on sliding windows

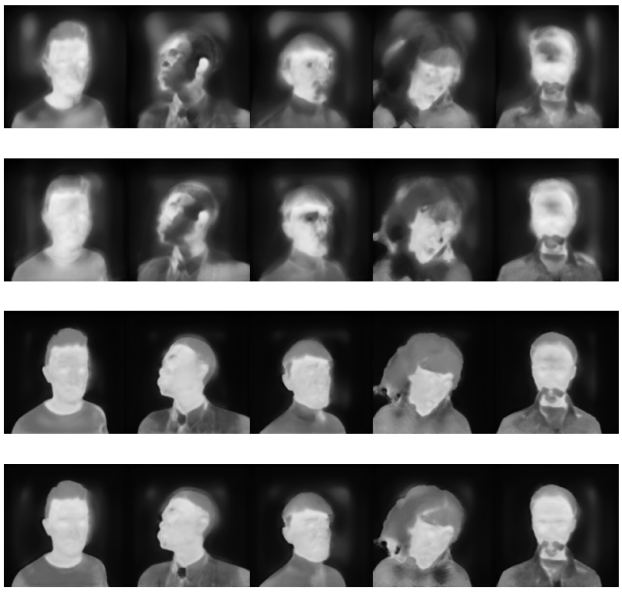

Figure 7. Regularization impact on images for 'warm' setup: No regularization ($\lambda_W = 0$, $\lambda_R = 0$), only patch-based ($\lambda_W = 0.01C$, $\lambda_R = 0$), only segmentation-based ($\lambda_W = 0$, $\lambda_R = 1$), and final model ($\lambda_W = 0.01C$, $\lambda_R = 1$) (top to bottom).

Figure 8. CHARLOTTE image samples with different resolutions, environmental conditions, and subjects.

of size 224×224 with a stride of 20. For our final landmark prediction, we pool all predictions and use the landmark with the smallest predicted standard deviation across all scales and all sliding windows. For training the landmarker T_ψ, we finetune a model that has been pre-trained

Metric	Method	Training Dataset	High	Low	Side	Front	Full
NME W/H ↓	GLL + RW ($\bar{\sigma} < \infty$) [21, 22]	FAKE	0.1055	0.2675	0.1241	0.2534	0.1887
	GLL + RW* ($\bar{\sigma} < \infty$)	FAKE	0.0933	0.2682	0.1312	0.2348	0.1824
	GLL + RW ($\bar{\sigma} < \infty$)	T-FAKE	0.0832	0.1334	0.0677	0.1503	0.1090
	GLL + RW ($\bar{\sigma} < \infty$)	FAKE + T-FAKE	0.0740	0.1346	0.0684	0.1420	0.1051

Table 4. Ablation results on CHARLOTTE dataset splits. Pre-processing with the pre-processing stack for RGB landmarkers is indicated by *. The confidence threshold has been set to infinity. RGB + Thermal (FAKE + T-FAKE) and Thermal Only (T-FAKE) models have been finetuned from the FAKE models.

on the original FAKE dataset for the sparse landmarker for the results in Table 4. During refinement, thermal images are used with a probability of 0.4 split with equal probability for 'cold' and 'warm' conditions. We finetune for 100 epochs with a learning rate of 0.0004, Adam optimizer with weight decay, batchsize of 512 and OneCycleLR scheduler. Our final model is trained for 4000 epochs, a learning rate of 0.001 on FAKE and T-FAKE (p=0.4).

Augmentations details. We use of spatial augmentations, including random shear, rotations, resizing, and cropping to allow the landmarker to learn a large variation of face orientations without the need for a dedicated face detector. Specifically, we use geometric augmentations which apply random rotations (up to ±45°) and random shear (±7°) to both the image and landmarks, preserving geometric consistency. In addition, a random resized cropping operation is performed, with the cropped region size scaling between 40% and 200% of the original image and an aspect ratio ranging between $\frac{3}{4}$ and $\frac{4}{3}$. Furthermore, we use photometric transformations and random Gaussian smoothing. The thermal images are randomly jittered to simulate temperature variations and thermal images are randomly inverted with a probability of 0.1. Additionally, random noise is applied with a probability of 0.2 to simulate sensor noise. See Figure 13 for image examples with augmentations applied.

2.3. Label Adaptation Implementation Details

Label Adaptation was trained for each method on the predictions on all detected faces on a random 1000 image CHARLOTTE split. We train a model T_ζ for 2000 epochs with a learning rate 0.002, OneCycleLR and Adam optimizer. As landmark augmentations, we apply random rotation up to 45° as well as random shearing during training. The label adaptation network is a five-layer perceptron with fully connected layers that takes the predicted landmarks together with the resize factor as input. The latter accounts for varying degrees of quantization at different image sizes in the CHARLOTTE ground truth. The label adaptation generally handles even outlier predictions but can also contain fail cases, (see Figure 10).

RGB Model Inference. We use two different pre-processing approaches to include landmarkers solely developed for RGB images into the evaluation. Firstly, gray-value images, where the temperature between 20° and 45° is normalized and, secondly the pre-processing stack proposed in [5]. The pre-processing stack consists of temperature clamping between 20°C and 45°C, unsharp masking with two sets of parameters with and without temperature inversion. The reported landmarks are the averages over all detected faces. This simple pre-processing stack is a simple method for boosting RGB landmarker performance for thermal images [5]. As a result, we can include RGB landmarkers as a baseline for thermal landmarking models.

2.4. Landmarking Ablation Study

To study the impact of our thermal data, we report the CHARLOTTE results of our landmarker trained with i) RGB images only, ii) finetuned with thermal images and iii) finetuned with both FAKE and T-FAKE. Again, we use label adaptation for all variations. Table 4 shows the results. Training with the T-FAKE dataset significantly improves the accuracy of thermal landmarking across all conditions. In addition, multimodal training with the FAKE and T-FAKE datasets leads to better thermal landmarking performance than finetuning only with the T-FAKE dataset.

2.5. Inference Ablation

We analyze the impact of our inference strategy, see Table 5. We compare inference on i.) the complete image scaled to 224×224 (whole image), ii.) followed by refinement on a bounding box computed from the predictions obtained with i.) and finally iii.) with the sliding window approach described in the main document. Method iii.) produces the best results over all images except for Charlotte low, where ii.) performs slightly better while at the same time being also suitable for real time estimation.

3. Large-Scale Visualization

For a large number of T-FAKE samples, we refer to Figures 11 and 12. Here, we simply use the first 128 images based on the numerical naming convention of the original FAKE dataset.

4. Thermal Semantic Segmentation Dataset

Note that by design detailed segmentation masks are available for all T-FAKE images. While the training of a se-

Figure 9. Results on examples from the **CHARLOTTE** dataset [1] with different RGB and thermal predictors and our models. For images without landmarks, no faces were detected. The performance of RGB methods can be greatly improved when the images are inverted or sharpened indicated by *. The first column shows the limitations in the CHARLOTTE ground truth: profile annotation convention for frontal views (1st row), quantization artefacts for low resolution images (4th row), translated annotations (last row).

Method	Inference Time (ms)	High	Low	Side	Front	Full
Sliding window ($\bar{\sigma} < \infty$)	88.73	0.0740	0.1346	0.0684	0.1420	0.1051
Refined bounding box ($\bar{\sigma} < \infty$)	10.41	0.0847	0.1329	0.0696	0.1494	0.1095
Whole image ($\bar{\sigma} < \infty$)	5.92	0.1091	0.1868	0.0842	0.2139	0.1490

Table 5. NME (W/H) (↓) for CHARLOTTE splits with different strategies for landmark computation. The final results are estimated with sliding windows similar to [22], however, we achieved comparable results when computing the landmarks on input images rescaled to 224×224. Here, we do not exclude high-uncertainty images and evaluate all images without failure rate, i.e., $\bar{\sigma} < \infty$. Average inference time per frame on the Full split has been benchmarked on a single NVIDIA H100 80GB GPU with a batch size of 1.

mantic segmentation model was out-of-scope for our work, we want to highlight the fact that our dataset can also be used for such training. The possibility of such segmentation training with synthetic data has already been demonstrated by Wood et al. [21].

5. Limitations

Thermalization. This work depends on the thermalization of the final renders in the FAKE dataset. The dataset contains very difficult lighting conditions and scene compositions that make it powerful to train landmarkers but also made the thermalization particularly challenging and could only be solved with advanced domain-adaptive semi-supervised regularization approaches. Despite a good perceptual result of the faces, some T-FAKE images can contain minor artifacts on clothing and on the background (e.g. see Figure 7, bottom second from right). Nevertheless, these artifacts remain limited. Moreover, background artifacts can easily be removed by choosing a suitable background based on the ground truth segmentation as implemented during our landmarker training, see Fig. 13. We merely used an MSE loss for our supervised training. Including an adversarial [10, 23] or a perceptual [11, 19] loss into our model might lead to perceptual improvement.

Dense Landmarks. In this work, we relied on the ground truth 70-point landmarks of the FAKE dataset [21] and dense Mediapipe [13] annotations. Training with the original 320- and 702-point landmarks could further boost accuracy and lead to an even denser landmarker. However, these landmarks are not publicly available. Furthermore, we only evaluate a mobilenet backbone for landmark detection without face tracking. Better performance for difficult poses and face variability could be achieved with denser models [22] and spatial normalization of face positions during training.

CHARLOTTE. The CHARLOTTE [1] dataset is among

the largest datasets with thermal recordings of faces that contain different levels of image quality as well as has a high variability in poses such as side profile pictures and tilting which makes the dataset ideal as a benchmark. However, the 2D annotation uses a convention where side profile images have a different number of landmarks than frontal faces. Furthermore, landmarks of low-resolution images are quantized and there are examples of shifted ground truth annotations (see Figure 9).

Label Adaptation. We retrain the label adaptation networks for each tested landmarker on its original predictions. Hence, for a given method, a high failure rate on CHARLOTTE means that fewer training images are available for that method. Also, poor predictions on some of the images such as profile images without failure produce a low quality of adapted landmarks (see Figure 10, TL54 dlib, column 3). It is important to note that the label adaptation does not use visual information from the benchmark dataset to translate landmarks. However, the same individuals were present in both the test and the training datasets which might allow networks to learn facial statistics of individuals. This might give an advantage to landmarkers with only a few landmarks, e.g., TFW [12] (see Figure 10, TFW). The good 68-point landmark performance of TFW on the CHARLOTTE dataset does not necessarily mean that this generalizes and the advantage of increasing the number of predicted landmarks has been demonstrated for RGB images [22].

References

[1] Roshanak Ashrafi, Mona Azarbayjani, and Hamed Tabkhi. Charlotte-ThermalFace: A fully annotated thermal infrared face dataset with various environmental conditions and distances. *Infrared Physics and Technology*, 124:104209, 2022. 1, 6

[2] Usman Cheema and Seungbin Moon. Sejong face database: A multi-modal disguise face database. *Computer Vision and Image Understanding*, 208:103218, 2021. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 1

[4] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *AISTATS*, pages 2681–2690. PMLR, 2019. 1

[5] Philipp Flotho, Mayur J Bhamborae, Tobias Grün, Carlos Trenado, David Thinnes, Dominik Limbach, and Daniel J Strauss. Multimodal data acquisition at sars-cov-2 drive through screening centers: Setup description and experiences in saarland, germany. *Journal of Biophotonics*, 14(8):e202000512, 2021. 2, 3, 5

[6] V. Fomin, J. Anmol, S. Desroziers, J. Kriss, and A. Tejani. High-level library to help with training neural networks in pytorch. https://github.com/pytorch/ignite, 2020. 3

[7] M. Godi, C. Joppi, A. Giachetti, F. Pellacini, and M. Cristani. Texel-att: Representing and classifying element-based textures by attributes. In *BMVC*, 2019. 4

[8] 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. *NIPS*, 30, 2017. 2

[9] Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, and Qingli Li. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *CVPR*, pages 18291–18300, 2022. 3

[10] 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. 3, 6

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *EECV*, pages 694–711. Springer, 2016. 6

[12] Askat Kuzdeuov, Dana Aubakirova, Darina Koishigarina, and Huseyin Atakan Varol. Tfw: Annotated thermal faces in the wild dataset. *IEEE TIFS*, 17:2084–2094, 2022. 7

[13] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *CVPR Workshops*, 2019. 6

[14] Osama Mazhar and Jens Kober. Random shadows and highlights: A new data augmentation method for extreme lighting conditions. *arXiv preprint arXiv:2101.05361*, 2021. 1

[15] Nithin Gopalakrishnan Nair and Vishal M Patel. T2v-ddpm: Thermal to visible face translation using denoising diffusion probabilistic models. In *International Conference on Automatic Face and Gesture Recognition*, pages 1–7. IEEE, 2023. 2

[16] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *ICLR*, 2018. 2

[17] Karen Panetta, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, Holly A Taylor, Arash Samani, et al. A comprehensive database for benchmarking imaging systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):509–520, 2018. 2

[18] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *EECV*, pages 319–345. Springer, 2020. 3

[19] Domenick D Poster, Shuowen Hu, Nathan J Short, Benjamin S Riggan, and Nasser M Nasrabadi. Visible-to-thermal transfer learning for facial landmark detection. *IEEE Access*, 9:52759–52772, 2021. 6

[20] Benjamin S Riggan, Nathaniel J Short, and Shuowen Hu. Thermal to visible synthesis of face images using multiple regions. In *WACV*, pages 30–38. IEEE, 2018. 2

[21] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *ICCV*, pages 3681–3691, 2021. 5, 6

[22] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljkovic, Tom Cash-

Figure 10. Label adaptation examples. For side profile views in the CHARLOTTE dataset only landmarks for visible parts of the face exist and additional positions on the forehead are marked. Label adaptation translates the three different landmark conventions we use for evaluation (bottom) into the CHARLOTTE convention (top).

man, and Julien Valentin. 3d face reconstruction with dense landmarks. In *ECCV*, 2022. 5, 6, 7

[23] Teng Zhang, Arnold Wiliem, Siqi Yang, and Brian Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *International Conference on Biometrics*, pages 174–181. IEEE, 2018. 2, 6

[24] 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. 3

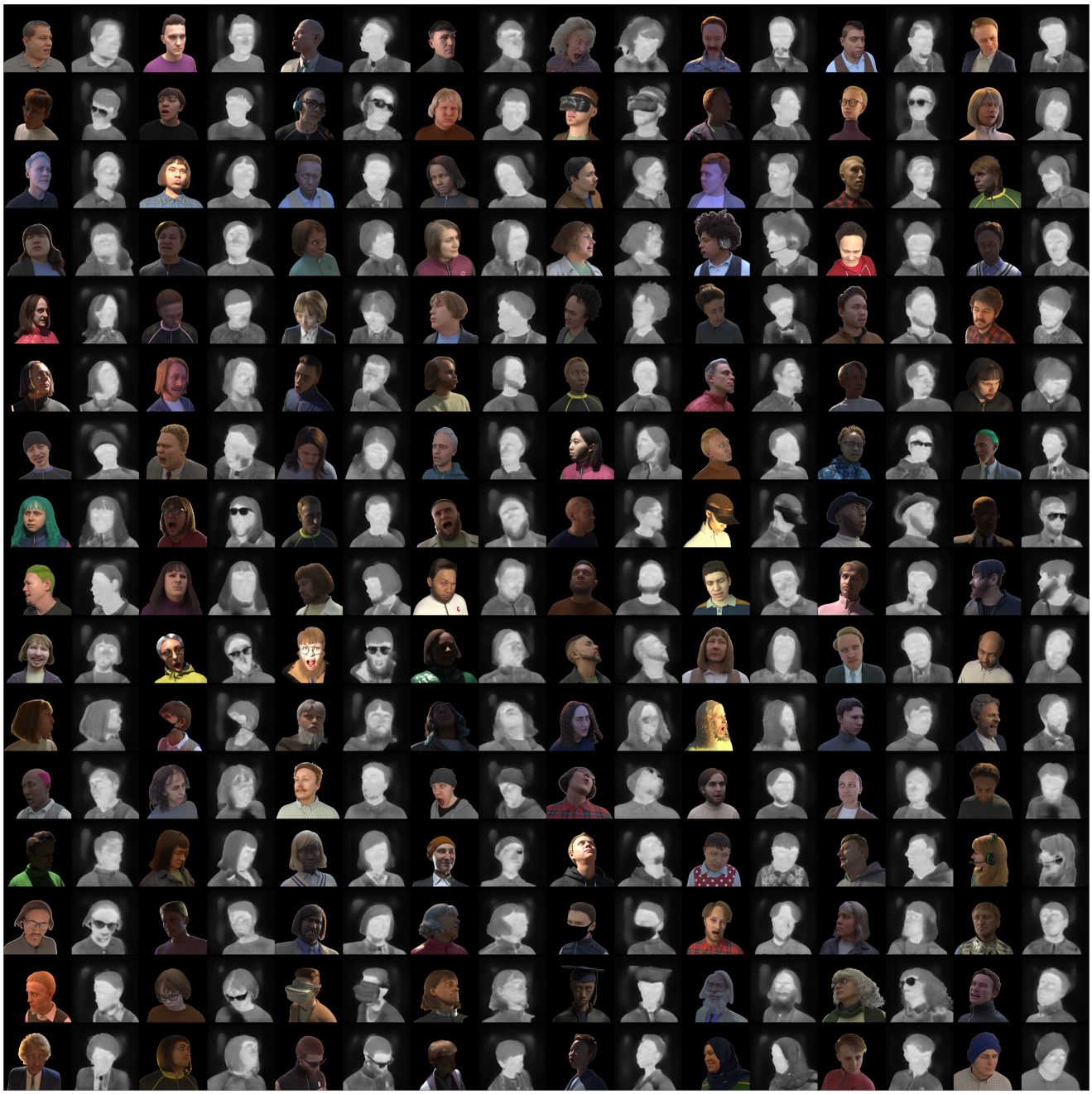

Figure 11. The first 128 FAKE (with removed background) and T-FAKE images with a random choice between the ‘cold’ and the ‘warm’ variant.

Figure 12. The sparse landmarks for the T-FAKE images in Fig. 11

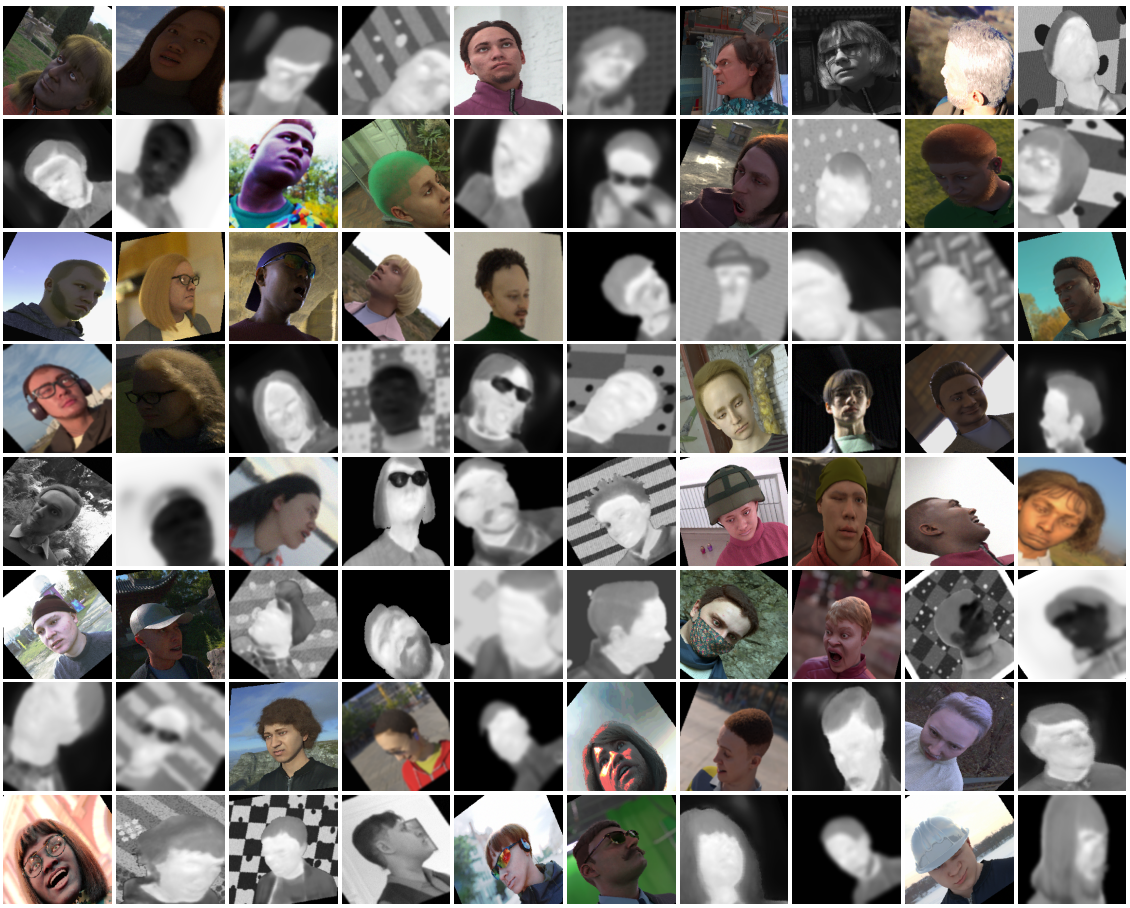

Figure 13. Example images from FAKE and T-FAKE with the augmentations applied during training.