Self-adversarial Multi-scale Contrastive Learning for Semantic Segmentation of Thermal Facial Images

Jitesh Joshi
jitesh.joshi.20@ucl.ac.uk
Nadia Bianchi-Berthouze
nadia.berthouze@ucl.ac.uk
Youngjun Cho*
youngjun.cho@ucl.ac.uk

Department of Computer Science
University College London
London, UK

Abstract

Reliable segmentation of thermal facial images in unconstrained settings such as thermal ambience and occlusions is challenging as facial features lack salience. Limited availability of datasets from such settings further makes it difficult to train segmentation networks. To address the challenge, we propose Self-Adversarial Multi-scale Contrastive Learning (SAM-CL) as a generic learning framework to train segmentation networks. SAM-CL framework constitutes SAM-CL loss function and a thermal image augmentation (TiAug) as a domain-specific augmentation technique to simulate unconstrained settings based upon existing datasets collected from controlled settings. We use the Thermal-Face-Database to demonstrate effectiveness of our approach. Experiments conducted on the existing segmentation networks- UNET, Attention-UNET, DeepLabV3 and HRNetv2 evidence the consistent performance gain from the SAM-CL framework. Further, we present a qualitative analysis with UBComfort and DeepBreath datasets to discuss how our proposed methods perform in handling unconstrained situations.

1 Introduction

Thermal infrared imaging of human skin enables remote extraction of physiological signals, along with monitoring of affective and psychological states [11]. For extraction of different physiological signals as well as affective states, studies indicate different regions such as the nostril region for breathing signal [38, 44], and perinasal region for stress and affective states [46]. Automated computational pipelines for processing thermal infrared images therefore require identification of regions of interest (ROIs) which can be either defined by a fixed bounding box or by an anatomical mask, with the later being more appropriate for reliable extraction of physiological signals [18]. The automated identification of an anatomical mask requires every pixel in a thermal image to be labelled for its respective anatomical region, which is a semantic segmentation task. In comparison with typical RGB or gray-scale facial images, thermal facial images possess much less prominent facial features. This is because

 * Corresponding Author

thermal images represent the temperature distribution over the skin surface, which is affected by dynamically varying physiological state as well as ambient temperature [13, 20]. In addition, variations over the thermal surface are very low and occlusion issues such as forehead hairs and eye-glasses further make it challenging to reliably segment the ROIs in unconstrained settings.

The state-of-the-art in semantic segmentation has been under continuous progression following the foundational work of Fully Convolutional Networks (FCN) [36]. The methodological contributions in the research of semantic segmentation can be broadly categorized into i) model architectures, ii) loss functions or learning strategies, and iii) data augmentation techniques. Significant development has been made towards the deep-learning architectures for semantic segmentation with some of the notable ones including UNET [45] and its variants [49], the family of DeepLab networks [6, 7, 8], HRNet [51], and the more recent transformer based approaches such as "HRNet + OCR + SegFix" [56]. Pretrained network backbones such as ResNet [23], Xception [14], and HRNet [51], have further accelerated the progress owing to the availability of large-scale bench-marking datasets [16, 24, 34].

On the other side, the widely used loss functions for semantic segmentation include softmax cross-entropy loss [15], DICE loss [50] and region mutual information (RMI) loss [60], among others. Furthermore, data-augmentation techniques applied to RGB images include basic image manipulations and deep learning approaches to augment the image representations [5, 47]. However, it is to be noted that the appearances of the objects in thermal imaging differ from that of RGB images. Not only is the data single channel, the basic properties such as transparency in RGB change to opacity in thermal images (e.g. clear glass). As thermal infrared wavelength is not transmissive for most of the objects, occlusions are observed more frequently. In addition, the variations in thermal ambient conditions result in varying appearances and can not be related to the brightness variations in RGB images.

The existing challenges in semantic segmentation of thermal images include the availability of large-scale bench-marking datasets, domain specific augmentation techniques to handle unconstrained real-world scenarios, and the studies validating the effectiveness of the architectures and loss functions, primarily proposed for color images. This work addresses the challenge of training segmentation network with limited dataset size using the proposed self-adversarial multi-scale contrastive learning (SAM-CL) framework (§3) comprising of the SAM-CL loss (§3.1.2) and the thermal image augmentation (TiAug) module (§3.2). The TiAug module serves as domain specific augmentation, while the SAM-CL loss provides enhanced supervision in learning inter-class separation and intra-class proximity in presence of adversarial-attacks by TiAug. We compare the performance of SAM-CL framework with the existing segmentation loss functions, supervised CL for segmentation [52] and GAN based approaches for segmentation [55, 59] in §4.1.2. Our contributions are summarised as:

- Self-Adversarial Multi-scale Contrastive-Learning (SAM-CL) framework for semantic segmentation with limited datasets in the thermal imaging domain, comprising of:
  - SAM-CL loss function that computes triplet loss for the predicted segmentation masks as anchor, ground-truth segmentation mask as positive and class-swapped mask as negative sample.
  - Thermal image augmentation (TiAug) module to transform the representation of thermal image acquired in controlled settings to the one acquired in unconstrained settings, contributing to the domain specific augmentation technique for thermal infrared imaging.
- Performance benchmark on existing Thermal Face Database [30, 31] for semantic segmentation of facial regions.

2 Related Work

2.1 Semantic Segmentation

Segmentation networks largely follow encoder-decoder schemes [36, 37, 40, 45, 62]. To learn cross-pixel dependencies, several models apply attention mechanisms [3] for semantic segmentation [19, 25, 26, 29, 41, 63]. Atrous convolutions along with pyramid pooling in *DeepLab* networks [6, 7, 8] enable learning of the multi-scale features. *HRNet*, as proposed in [51], shows performance gain on the semantic segmentation task, among other tasks, by making use of high-resolution and multilevel representations. Furthermore, the more recent development using transformer networks such as *HRNet + OCR + SegFix* has achieved competitive performances on multiple benchmarking datasets [56]. While increased network complexity and deeper layers prove effective for training models with large-scale RGB datasets, it is challenging to benefit from the same with a limited dataset size as typically observed in the case of thermal infrared imaging.

Existing semantic segmentation approaches for thermal images are not equipped to handle real-world scenarios, such as occlusion and varying thermal ambient conditions. One earlier study on occlusion removal in thermal imaging [53] proposes a modelling-based method using kernel principal component analysis for removing a specific occluding object (eye-glasses). This method requires the use of a registered color image to reconstruct the occluded thermal image, limiting its generalizability unless a large-scale dataset with pair of color-images and thermal-images is available. Large scale datasets allow capturing diverse representations, though acquiring a large scale dataset with thermal imaging and performing pixel wise annotations for semantic segmentation remains impractical. Furthermore, the thermal imaging datasets that are currently available with facial images have been acquired in highly controlled settings [1, 30, 31, 33]. It is therefore required to review the data augmentation techniques that can allow achieving robust performance in real-world scenarios.

2.2 Image Augmentation Techniques

The commonly used augmentation techniques include geometric transformations as well as learning or modelling based methods [47]. While geometric transformations are relevant for thermal images, augmentation techniques pertaining to variations in brightness and contrast in RGB images cannot be directly mapped to thermal images. Among the learning based methods, GAN [21] and self-adversarial training (SAT) [22] have shown promising performance. SimGAN as proposed in [48] utilizes simulator generated images and a GAN to synthesize realistic augmented eye images. This method relies on the effectiveness of a simulator in synthesizing images, which may not be generalization for different scenes and image modalities. YOLOv4 [4] showed the effectiveness of SAT based augmentation technique called Fast Gradient Sign Method (FGSM) [22] in which a network updates an original image instead of the network weights in one forward pass, and this altered image is then used as an adversarial attack to improve the robustness of the trained model. A more recent work on localization of image forgery [64] also highlights the usefulness of FGSM based self-adversarial attacks in augmenting the data. While existing SAT approaches increase the robustness of the model for subtle changes in an image, they are insufficient in modelling range of real-world scenarios. Our proposed method (§3.2) is inspired from the idea of modelling based dynamic generation of plausible variations in thermal images for adversarial attacks to enhance the robustness of the trained model.

2.3 Loss Functions or Learning Strategies

In addition to segmentation network and augmentation techniques, loss function or learning strategy plays a crucial role in achieving the higher performance. For semantic segmentation tasks, cross-entropy loss [15] and weighted cross-entropy loss functions have been widely used [2, 28]. In cases of unequal class distribution, due to imbalanced distribution of pixels between semantic classes, focal loss [35] and DICE loss [50] have been reported to show better performance. In a more recent development, researchers proposed a loss function based on mutual information between pixels and semantic regions [60], and showed substantial improvements on benchmarking datasets. Learning strategies such as GAN [10, 17, 21, 55, 57, 59] and CL [52, 58, 61] have also shown to be effective for the segmentation task. A recent study using GAN [59] proposes multi-class segmentation approach, though, it is limited to handle only the occluding objects learnt at the training time. While the GAN requires larger datasets, the corner stone for the success of CL approaches is the availability of pretrained models, which requires large scale datasets of the same imaging domain, though with a flexibility of different upstream computer vision tasks. Our work takes inspiration from the GAN as well as the CL, however unlike the critic network in GAN, the auxiliary network in our approach does not compete with the generator (segmentation network), and unlike the use of feature space for sampling anchors in CL, our approach uses predicted segmentation masks or logits as anchors.

3 Proposed Method: Self-Adversarial Multi-scale Contrastive Learning (SAM-CL)

In this section, we first describe our proposed Self-Adversarial Multi-scale Contrastive Learning (SAM-CL) framework for semantic segmentation. Though this framework can be generically applied to segmentation tasks, the key objective in this work is to train a segmentation network on thermal facial images for: i) segmentation of facial regions including eyes, eyebrows, nose, mouth and chin area and, ii) resilience to varying thermal ambient conditions as well as occlusions in unconstrained settings. As this needs to be achieved without the availability of thermal images acquired in such scenarios, we propose the SAM-CL framework comprising of the SAM-CL loss function and the thermal image augmentation (TiAug) module. Figure 1 depicts an overview of the SAM-CL framework which is required only during the training, resulting in no overhead of computation during the inference.

3.1 Loss Function

3.1.1 Preliminaries

For a semantic segmentation task, the segmentation network SEG learns a function $f_{SEG}(I)$ that maps input image I to the ground-truth mask Y, that specifies a semantic class $c \in C$ for every pixel $i \in I$. The limitation of the most commonly used pixel-wise segmentation loss functions such as cross-entropy loss, is their inability to capture relationships between pixels. To address this limitation, a recent work proposed mutual information based loss function [60] that combines cross-entropy and structural similarity loss. Furthermore, to learn the relationship between pixels of multiple images and to supervise the representations within pixel-embedding, supervised contrastive loss for semantic segmentation is proposed in [52]:

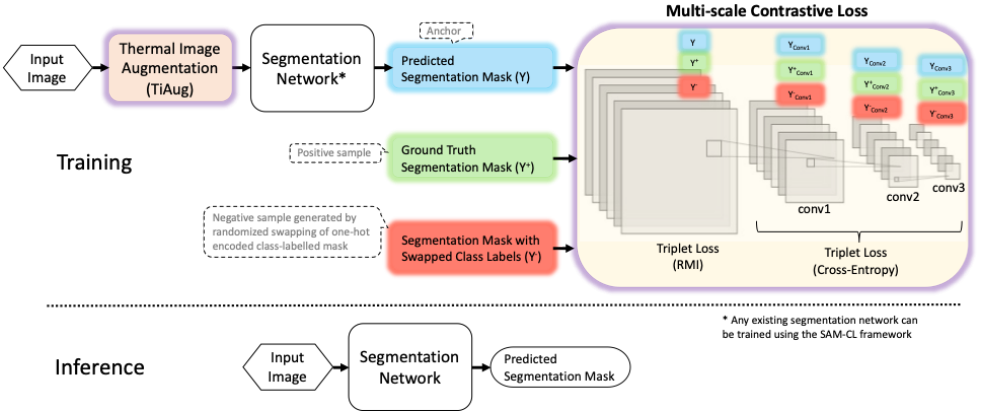

Figure 1: Proposed SAM-CL framework for Semantic Segmentation. In SAM-CL framework, multi-scale contrastive-loss is computed by the auxiliary network. TiAug add occlusions as well as range of plausible variations representing real-world scenarios to thermal images, providing adversarial attacks for training the segmentation network.

$$\mathcal{L}_y^{NCE} = \frac{1}{|P_y|} \sum_{y^+ \in P_y} -log \frac{\exp(y \cdot y^+/\tau)}{\exp(y \cdot y^+/\tau) + \sum_{y^- \in N_y} \exp(y \cdot y^-/\tau)} \tag{1}$$

where P_y and N_y are positive and negative samples of pixel-embedding, belonging to classes dissimilar to that of the anchor pixel y. This learning strategy is very effective in maximizing inter-class separation, while minimizing intra-class distance within pixel-embedding, specially for large-scale datasets. However, while training a segmentation network without pretrained weights and with a limited dataset size, representations within pixel-embeddings and the feature maps of earlier layers remain highly transient thereby making the use of contrastive learning challenging. To address this, we resort to a learning strategy that does not require the use of pixel-embedding, while being effective in maximizing inter-class separation and minimizing intra-class distance.

3.1.2 SAM-CL Loss Function

In a one-hot encoded ground-truth segmentation mask (Y_{oh}^+), each channel represents a binary mask for the respective classes. Class swapped mask (Y_{oh}^-) is generated by randomized swapping of channels of the (Y_{oh}^+) with a constraint that no channels of Y_{oh}^+ and Y_{oh}^- match. With logits or one-hot predicted mask Y_{oh} representing an anchor, Y_{oh}^+ and Y_{oh}^- representing positive and negative samples respectively, the first triplet loss is computed as shown in Equation (2):

$$\mathcal{L}_{s0}(Y_{oh}, Y_{oh}^+, Y_{oh}^-) = \max\{d(Y_{oh}, Y_{oh}^+) - d(Y_{oh}, Y_{oh}^-) + margin, 0\} \tag{2}$$

Equation (2) allows learning inter-class separation as well as intra-class proximity without requiring to compute the contrastive loss with pixel-embedding. As Y_{oh}^- preserves spatial features at mask-level, the optimization results in effective inter-class separation of the spatial features. Y_{oh}, Y_{oh}^+, and Y_{oh}^- are passed through a 4-layered auxiliary network in three different

forward passes to compute the feature maps y_{Conv1}, y_{Conv2} and y_{Conv3}; $(y = Y_{oh}, Y^{+}_{oh}, Y^{-}_{oh})$ for each layer. Down-scaling of 2 is applied at each layer with the number of channels in every layer held constant and equal to the number of classes, to match with the one-hot encoded class-labels. The final SAM-CL loss function as formulated in Equation (3), therefore offers supervision to maximize inter-class separation at multiple-scales. To compute the distance function $d(x,y)$ as mentioned in Equation (2), we use RMI [60] on logits, and cross-entropy loss for down-convolved feature-maps as shown in Figure 1.

$$\mathcal{L}_{SAM-CL} = \mathcal{L}_{s0}(Y_{oh}, Y^{+}_{oh}, Y^{-}_{oh}) + \mathcal{L}_{s1}(Y_{Conv1}, Y^{+}_{Conv1}, Y^{-}_{Conv1}) + \mathcal{L}_{s2}(Y_{Conv2}, Y^{+}_{Conv2}, Y^{-}_{Conv2}) + \mathcal{L}_{s3}(Y_{Conv3}, Y^{+}_{Conv3}, Y^{-}_{Conv3}) \quad (3)$$

3.2 Thermal Image Augmentation (TiAug) Module

The thermal image augmentation module (TiAug) transforms a thermal image acquired in controlled settings into an image resembling the one acquired in unconstrained ambient settings. This is achieved by first adding synthesized objects with diverse characteristics in an occluding as well as a non-occluding manner. Secondly, to every pixel, a random temperature value is added as thermal noise. The maximum magnitude of the noise is set as per the noise equivalent temperature difference (NETD), a sensitivity parameter of thermal infrared imaging camera, that provides a minimum value of temperature difference that can be sensed reliably by the camera. While a high-sensitive thermal camera has lower magnitude of NETD, it is higher for the low-sensitive mobile thermal imaging camera. TiAug sets the maximum NETD value ($Th^{max}_{NETD} = 0.1°C$) considering the widely used mobile thermal imaging camera. Figure 2 provides an overview of the sub-modules, list of parameters as well as illustrative samples, while the image transformation is expressed in Equation (4).

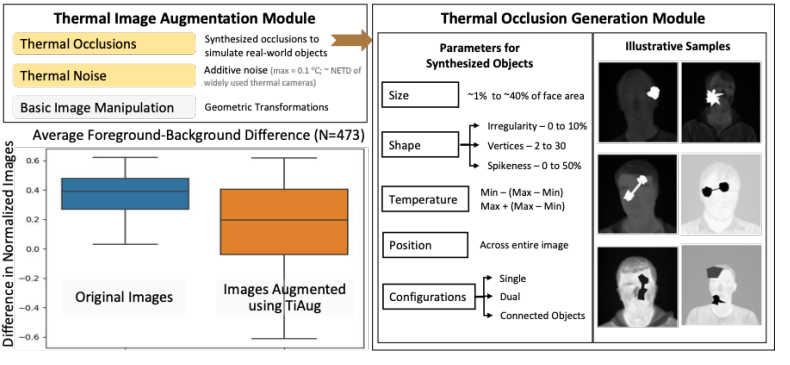

Figure 2: Thermal Image Augmentation (TiAug) Module. TiAug module comprises of the modules for thermal data specific occlusion generation as well as additive thermal noise along with the commonly used geometric transformations.

$$I_{aug} = f_{occ}(I_{org}, g(\theta)) + \eta \quad (4)$$

I_{org} is the original image; the augmented image I_{aug} is a resultant of the transformation f_{occ} and a matrix with additive thermal noise η $(0 < \eta < Th^{max}_{NETD})$. The θ in f_{occ} represents a set of parameters used for synthesizing objects as listed with the corresponding ranges in Figure 2. The function $g(\theta)$ uniquely combines the values of each parameter from the respective range, allowing the synthesis of endless new representations. f_{occ} replaces the

specific pixels of I_{org} with the synthesized objects which may or may not occlude the facial regions, resulting in altered histogram of thermal images which typically exhibit bimodal characteristics when acquired in controlled settings. In a bimodal histogram distribution, a peak at a lower temperature value depicts background and a second peak at a higher temperature corresponds to the facial regions. Several segmentation algorithms assume bimodal histogram distribution for an automated segmentation [27], which affects their performance in unconstrained settings.

Real-world scenarios may include objects at temperatures higher than the facial regions (e.g., sun, hot beverages), as well as objects at lower temperatures. In such real-world scenarios, histogram distribution does not depict bimodal characteristics. To simulate the real-world variations in histogram distribution pattern, TiAug adds synthesized objects both at the higher and the lower temperature than that of the facial regions, which in-turn prevents the deep-learning network from over-fitting to the bimodal-distribution. I_{aug} thus obtained, represents real-world scenarios both in terms of spatial characteristics of ambient objects as well as the histogram distribution of temperature values. I_{aug} is further normalized and passed to the segmentation network as an input. While the examples in the Figure 2 show changes in spatial characteristics, the box-plot analysis signifies effectiveness of TiAug in altering average foreground (facial-regions) temperature and average background temperature in normalized images. Furthermore, TiAug applies following geometric transformations: horizontal flip, vertical flip, rotation, Gaussian blur, and resizing (0.5X to 2X).

4 Experiments

We perform experiments to compare the proposed SAM-CL framework with existing loss functions and learning strategies, using the segmentation networks including U-NET [45], Attention UNET [41], DeepLabV3 [9] and HRNet [51]. Our code, which is available here, uses PyTorch [43] and is built upon a prior work on contrastive learning for semantic segmentation [52]. Given a lack of benchmark segmentation performance reports on thermal facial datasets as well as pretrained models, we implement and train the aforementioned prior-art segmentation networks with Xavier uniform initialization. We use a batch size of 16 along with an SGD optimiser with a weight decay of 1e-8 and betas set to 0.9 and 0.999.

For the experiments, it is required to use datasets of raw thermal matrices (ie. absolute temperature value assigned to each pixel). With this, we have carefully chosen three datasets of thermal facial images: Thermal Face Database [30, 32], UBComfort dataset [42], and DeepBreath dataset [12, 13]. For training and quantitative evaluation (§4.2), we mainly use the Thermal Face Database [30, 32] as it comprises of ground-truth labels for images acquired in controlled setting, while the later two datasets collected from uncontrained settings are used for qualitative analysis (§4.2) to demonstrate the effectiveness of SAM-CL framework in such settings.

4.1 Quantitative Evaluation on Thermal Face Database

4.1.1 Dataset Description

Thermal Face Database [30, 32] comprises of 2935 images of 90 individuals with manually annotated 68 facial landmark points. We derive segmentation masks from the landmarks points by labelling all the pixels enclosed within the boundary formed by the landmark points

for each anatomical region including chin, mouth, nose, eyes and eye-brows. We split the data into training (85%) and validation (15%) sets based on subject ids, to evaluate the generalisation of the model based on the performance obtained on the validation set.

4.1.2 Results

To investigate the generalizability of the SAM-CL framework, we train UNET [45], Attention UNET [41], DeepLabV3 [9] and HRNet [51] segmentation networks. In all the experiments, augmentation with basic geometric transformations including horizontal flip, vertical flip, rotation, Gaussian blur, and resizing (0.5X to 2X), is uniformly applied. The loss functions used for bench-marking includes weighted binary cross-entropy loss (BCE), DICE loss, and region mutual information (RMI) loss [60]. As SAM-CL framework relates to CL and GAN, we additionally compare performance with SegAN [54, 55], SegGAN [59], along with a recent work on the supervised CL applied to semantic segmentation [52]. Table 1 shows the comparison of performances with percentage mean IoU metric.

Table 1: Performance Evaluation of SAM-CL Framework

Segmentation Network	Learning Strategy (Loss Function)	mIoU (%)	Segmentation Network	Learning Strategy (Loss Function)	mIoU (%)
UNET [■]	Pixel-wise Segmentation (BCE)	67.64	Attention UNET [■]	Pixel-wise Segmentation (BCE)	66.61
	Pixel-wise Segmentation (DICE)	75.00		Pixel-wise Segmentation (DICE)	75.14
	GAN (SegAN) [■]	76.79		GAN (SegAN) [■]	76.75
	GAN (SegGAN) [■]	75.50		GAN (SegGAN) [■]	76.24
	RMI [■]	81.35		RMI [■]	81.39
	ContrastiveSeg [■]	81.24		ContrastiveSeg [■]	81.50
	SAM-CL (Ours)	**82.11 (+0.76)**		**SAM-CL (Ours)**	**82.85 (+1.35)**
DeepLabV3+ResNet101 [■, ■]	RMI [■]	75.85	HRNetV2-W48 [■]	RMI [■]	78.46
	ContrastiveSeg [■]	74.45		ContrastiveSeg [■]	78.36
	SAM-CL (Ours)	**79.29 (+3.44)**		**SAM-CL (Ours)**	**78.97 (+0.61)**

We observe consistent performance gain with the use of DICE loss when compared against BCE loss. GAN based learning strategy [55, 59] is found effective for UNET and Attention UNET, while the performance drops for DeepLabV3 network when comparing against the respective performance with the DICE loss. Consistent performance improvement from DICE loss function as well as GAN based learning strategy is evident for the models trained with RMI loss function. The performance gains across all the segmentation networks can be noted when deploying SAM-CL framework.

We performed an ablation study to examine the individual contribution of the TiAug module and the SAM-CL loss function. For ablation study, RMI loss function was used uniformly across all the experiments. From Table 2, we observe performance gains from the baseline for all the segmentation networks when the data is augmented using the TiAug module. Similarly additional performance gains are observed when the SAM-CL loss function is used for optimization. The combination of the SAM-CL loss function and the TiAug module outperforms as TiAug presents a network with the adversaries such as occlusions and varying ambient temperature levels in the input thermal images, while the SAM-CL loss maximizes the inter-class separation using the class-swapped negative sample Y^- and its down-scaled representations in the auxiliary network.

4.2 Qualitative Analysis

The Thermal Face Database does not include real-world occlusions, and is acquired in highly controlled laboratory settings. We therefore extend the evaluation with qualitative analysis on the datasets acquired in unconstrained settings as depicted in Figure 3. The UBComfort

Table 2: Ablation Study for TiAug and SAM-CL Loss

Segmentation Network	mIoU (%) Performance		
	RMI	RMI + TiAug	RMI + TiAug + SAM-CL
UNET [8]	81.36	81.91	**82.11**
Attention UNET [8]	81.39	82.29	**82.85**
HRNetV2-W48 [8]	78.13	78.87	**78.97**
DeepLabV3+ResNet101 [8, 8]	75.85	78.07	**78.12**
DeepLabV3+Xception [8, 8]	76.55	77.31	**77.85**

dataset [42] was acquired from in-the-wild car users with varying thermal ambient conditions using a high-resolution thermal camera, whereas the DeepBreath dataset [12] was acquired from participants exposed to different induced stress levels using a low-resolution mobile thermography camera. These datasets did not have ground-truth labelmasks, limiting their use in running quantitative evaluation studies.

The thermal images in the first row of the Figure 3 present ground-truth labelmask overlaid with color-coded class-labels. The histogram plot on the right-top (in Figure 3), shows the temperature distribution across the entire set of thermal images in the Thermal Face database, highlighting the highly controlled laboratory settings. We further identify a few samples within the Thermal Face Database in which hairs occlude a small part of thermal image. The superiority of the the model trained jointly with the SAM-CL loss function and the TiAug module in reliably handling forehead hairs as occlusions is evidenced by the segmentation outcome as shown in row-2 of Figure 3. In the following row, we present samples generated by the TiAug module, which after min-max normalization, appear significantly different, as would be the case when objects that are either too hot or too cold appear in an image. The state-of-the-art (SOTA) model fails in these scenarios (see row-3, Figure 3) as it has not been trained with such variations in the appearance.

The thermal images of individuals without and with eye-glasses, seated in a car [42] are shown in row-4 and 5 of Figure 3 respectively. Though the training set does not include any images with eye-glasses, the model trained with our proposed method shows resilience towards performing reliable segmentation in the presence of eye-glasses, while the SOTA sub-performs. Similary, the thermal images of individuals performing cognitive tasks [12], without and with eye-glasses, are shown in row-6 and 7 of Figure 3 respectively. While SOTA sub-performs on both the cases, SAM-CL framework shows reliable performance. It can be noted that thermal images in DeepBreath dataset [12] are acquired using mobile thermal camera (FLIR One), highlighting the robustness of the model trained using SAM-CL framework across for different thermal camera specifications.

5 Conclusion

The SAM-CL framework, comprising of the SAM-CL loss function and the TiAug module, is shown to be effective in training segmentation networks with limited dataset sizes. TiAug module presents the segmentation network with adversaries (e.g. occluded images), which results in a portion of predicted logits to overlap with the synthesized negative sample (from class-swapped labelmask). This overlap of logits with negative sample results in a higher loss for the incorrect label predicted by the segmentation network in the corresponding region. This explains the effectiveness of the SAM-CL loss function in conjunction with the TiAug module, in offering consistent performance gain across different segmentation networks. SAM-CL loss function can be extended to other imaging modalities to train segmentation

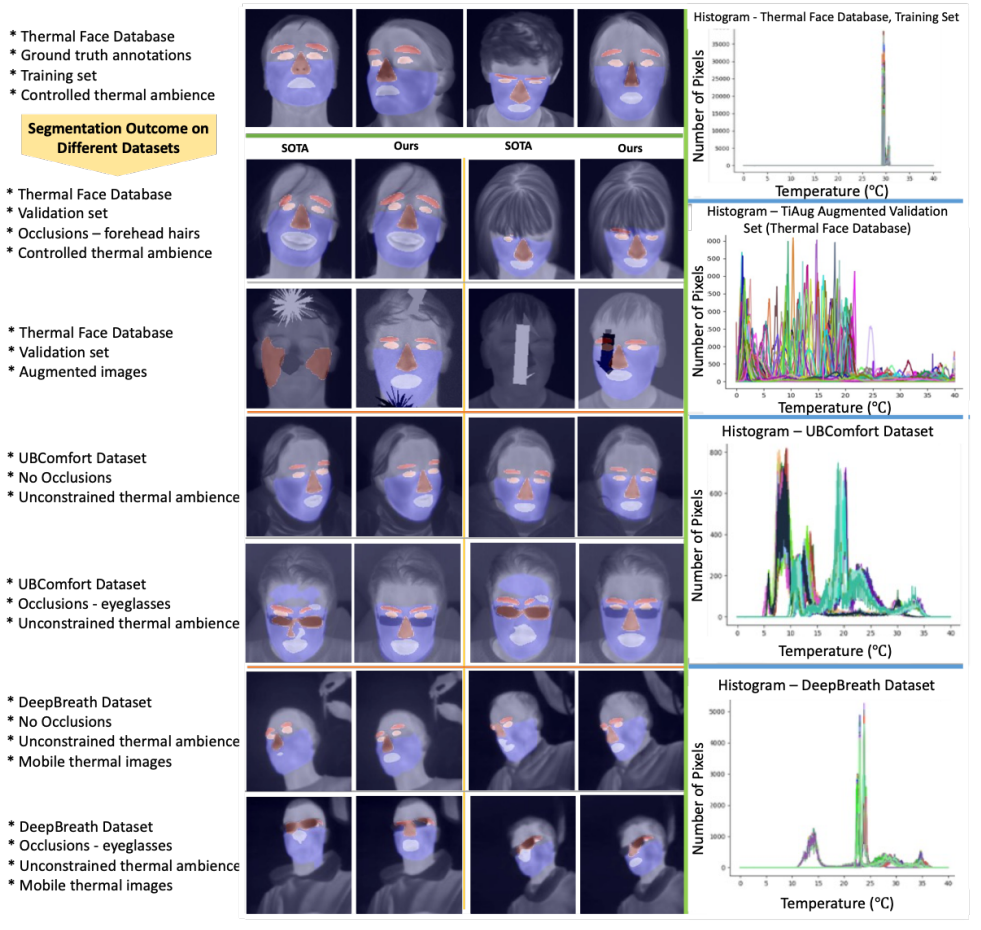

Figure 3: Qualitative Performance Analysis on Different Datasets. Both our work and SOTA use Attention UNET model trained without and with SAM-CL framework respectively. Please refer to §4.2 for the discussion on qualitative performance analysis.

network with limited dataset, with the addition of appropriate augmentation technique that adds challenging artifacts or adversaries.

While the performance of SAM-CL loss function relies on the TiAug module, the later can be utilized independently to transform the thermal images collected in controlled laboratory environment to the thermal images resembling several of the common real-world scenarios. For this transformation, the TiAug considers a range of plausible ambient temperature, basic geometric properties of common occluding objects as well as noise specification of widely used thermal cameras. The TiAug module can be deployed in different computer vision tasks utilizing thermal images including classification, object detection, instance and panoptic segmentation to enable training of the respective deep-learning networks to handle common real-world scenarios, without explicitly requiring a thermal dataset to be acquired in such scenarios.

References

[1] Madina Abdrakhmanova, Askat Kuzdeuov, Sheikh Jarju, Yerbolat Khassanov, Michael Lewis, and Huseyin Atakan Varol. SpeakingFaces: A Large-Scale Multimodal Dataset of Voice Commands with Visual and Thermal Video Streams. *Sensors*, 21(10):3465, January 2021. doi: 10.3390/s21103465. URL https://www.mdpi.com/1424-8220/21/10/3465. ZSCC: 0000004 Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54(1): 137–178, January 2021. ISSN 1573-7462. doi: 10.1007/s10462-020-09854-1. URL https://doi.org/10.1007/s10462-020-09854-1. ZSCC: 0000115.

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple Object Recognition with Visual Attention. *arXiv:1412.7755 [cs]*, April 2015. URL http://arxiv.org/abs/1412.7755. arXiv: 1412.7755.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *http://arxiv.org/abs/2004.10934*, April 2020. URL http://arxiv.org/abs/2004.10934.

[5] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2): 125, February 2020. ISSN 2078-2489. doi: 10.3390/info11020125. URL https://www.mdpi.com/2078-2489/11/2/125. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv:1606.00915 [cs]*, May 2017. URL http://arxiv.org/abs/1606.00915. arXiv: 1606.00915.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv:1706.05587 [cs]*, December 2017. URL http://arxiv.org/abs/1706.05587. ZSCC: 0003216 arXiv: 1706.05587.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, April 2018. ISSN 1939-3539. doi: 10.1109/TPAMI.2017.2699184. ZSCC: 0000003 Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. pages 801–818, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Liang-Chieh_Chen_Encoder-Decoder_with_Atrous_ECCV_2018_paper.html.

[10] Anoop Cherian and Alan Sullivan. Sem-GAN: Semantically-Consistent Image-to-Image Translation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1797–1806, January 2019. doi: 10.1109/WACV.2019.00196. ZSCC: 0000029 ISSN: 1550-5790.

[11] Youngjun Cho and Nadia Bianchi-Berthouze. Physiological and Affective Computing through Thermal Imaging: A Survey. *arXiv:1908.10307 [cs]*, August 2019. URL http://arxiv.org/abs/1908.10307. arXiv: 1908.10307.

[12] Youngjun Cho, Nadia Bianchi-Berthouze, and Simon J. Julier. DeepBreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 456–463, October 2017. doi: 10.1109/ACII.2017.8273639. ISSN: 2156-8111.

[13] Youngjun Cho, Simon J. Julier, Nicolai Marquardt, and Nadia Bianchi-Berthouze. Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging. *Biomedical Optics Express*, 8 (10):4480–4503, October 2017. ISSN 2156-7085. doi: 10.1364/BOE.8.004480. URL https://www.osapublishing.org/boe/abstract.cfm?uri=boe-8-10-4480. Number: 10 Publisher: Optical Society of America.

[14] Francois Chollet. Xception: Deep Learning With Depthwise Separable Convolutions. pages 1251–1258, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html.

[15] Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, 134(1):19–67, February 2005. ISSN 1572-9338. doi: 10.1007/s10479-005-5724-z. URL https://doi.org/10.1007/s10479-005-5724-z. ZSCC: 0002154.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

[17] Xue Dong, Yang Lei, Tonghe Wang, Matthew Thomas, Leonardo Tang, Walter J. Curran, Tian Liu, and Xiaofeng Yang. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. *Medical Physics*, 46(5):2157–2168, 2019. ISSN 2473-4209. doi: 10.1002/mp.13458. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13458. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13458.

[18] A. Duarte, L. Carrão, M. Espanha, T. Viana, D. Freitas, P. Bártolo, P. Faria, and H. A. Almeida. Segmentation Algorithms for Thermal Images. *Procedia Technology*, 16:1560–1569, January 2014. ISSN 2212-0173. doi: 10.1016/j.protcy.2014.10.178. URL http://www.sciencedirect.com/science/article/pii/S2212017314004058. ZSCC: 0000042.

[19] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual Attention Network for Scene Segmentation. pages 3146–3154, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Fu_Dual_Attention_Network_for_Scene_Segmentation_CVPR_2019_paper.html. ZSCC: 0001493.

[20] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: a survey. *Machine Vision and Applications*, 25(1):245–262, January 2014. ISSN 1432-1769. doi: 10.1007/s00138-013-0570-5. URL https://doi.org/10.1007/s00138-013-0570-5. ZSCC: 0000492.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html. ZSCC: 0035398.

[22] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, March 2015. URL http://arxiv.org/abs/1412.6572. arXiv:1412.6572 [cs, stat].

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. pages 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html. ZSCC: 0109350.

[24] Derek Hoiem, Santosh K. Divvala, and James H. Hays. Pascal VOC 2008 Challenge, 2009. ZSCC: 0000034.

[25] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. pages 7132–7141, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper. ZSCC: 0008551.

[26] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. CCNet: Criss-Cross Attention for Semantic Segmentation. pages 603–612, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Huang_CCNet_Criss-Cross_Attention_for_Semantic_Segmentation_ICCV_2019_paper.html. ZSCC: 0000643.

[27] Akshay Isalkar and K. Manikandan. Analysis of Image Segmentation Algorithms for Infrared Images. In Amit Dhawan, Vijay Shanker Tripathi, Karm Veer Arya, and Kshirasagar Naik, editors, *Recent Trends in Electronics and Communication*, Lecture Notes in Electrical Engineering, pages 639–646, Singapore, 2022. Springer. ISBN 9789811627613. doi: 10.1007/978-981-16-2761-3_57.

[28] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, October 2020. doi: 10.1109/CIBCB48159.2020.9277638. ZSCC: 0000165.

[29] Qiangguo Jin, Zhaopeng Meng, Changming Sun, Hui Cui, and Ran Su. RA-UNet: A Hybrid Deep Attention-Aware Network to Extract Liver and Tumor in CT Scans. *Frontiers in Bioengineering and Biotechnology*, 0, 2020. ISSN 2296-4185. doi: 10.3389/fbioe.2020.605132. URL https://www.frontiersin.org/articles/10.3389/fbioe.2020.605132/full. ZSCC: 0000073 Publisher: Frontiers.

[30] M. Kopaczka, R. Kolk, and D. Merhof. A fully annotated thermal face database and its application for thermal facial expression recognition. In *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6, May 2018. doi: 10.1109/I2MTC.2018.8409768. ZSCC: 0000021.

[31] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof. A Thermal Infrared Face Database With Facial Landmarks and Emotion Labels. *IEEE Transactions on Instrumentation and Measurement*, 68(5): 1389–1401, May 2019. ISSN 1557-9662. doi: 10.1109/TIM.2018.2884364. ZSCC: NoCitationData[s0] Conference Name: IEEE Transactions on Instrumentation and Measurement.

[32] Marcin Kopaczka, Lukas Breuer, Justus Schock, and Dorit Merhof. A Modular System for Detection, Tracking and Analysis of Human Faces in Thermal Infrared Recordings. *Sensors*, 19(19):4135, January 2019. doi: 10.3390/s19194135. URL https://www.mdpi.com/1424-8220/19/19/4135. ZSCC: 0000003 Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

[33] M. Kowalski and A. Grudzień. High-resolution thermal face dataset for face and expression recognition. *Metrology and Measurement Systems*, Vol. 25(nr 2), 2018. ISSN 0860-8229. doi: 10.24425/119566. URL http://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-d00dadcc-8eaf-4883-9b6e-a1a66fd44ba6. ZSCC: 0000006.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. pages 2980–2988, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html.

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. pages 3431–3440, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html.

[37] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, October 2016. doi: 10.1109/3DV.2016.79. ZSCC: 0004038.

[38] R. Murthy and I. Pavlidis. Noncontact measurement of breathing function. *IEEE Engineering in Medicine and Biology Magazine*, 25(3):57–67, May 2006. ISSN 1937-4186. doi: 10.1109/MEMB.2006.1636352. Conference Name: IEEE Engineering in Medicine and Biology Magazine.

[39] David Müller, Andreas Ehlen, and Bernd Valeske. Convolutional Neural Networks for Semantic Segmentation as a Tool for Multiclass Face Analysis in Thermal Infrared. *Journal of Nondestructive Evaluation*, 40(1):9, January 2021. ISSN 1573-4862. doi: 10.1007/s10921-020-00740-y. URL https://doi.org/10.1007/s10921-020-00740-y. ZSCC: 0000004.

[40] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning Deconvolution Network for Semantic Segmentation. pages 1520–1528, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Noh_Learning_Deconvolution_Network_ICCV_2015_paper.html.

[41] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv:1804.03999 [cs]*, May 2018. URL http://arxiv.org/abs/1804.03999. ZSCC: 0000934 arXiv: 1804.03999.

[42] Temitayo Olugbade, Youngjun Cho, Zak Morgan, Mohamed Abd El Ghani, and Nadia Bianchi-Berthouze. Toward Intelligent Car Comfort Sensing: New Dataset and Analysis of Annotated Physiological Metrics. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, September 2021. doi: 10.1109/ACII52823.2021.9597393. ZSCC: 0000000 ISSN: 2156-8111.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html. ZSCC: NoCitationData[s0].

[44] Carina Barbosa Pereira, Xinchi Yu, Michael Czaplik, Rolf Rossaint, Vladimir Blazek, and Steffen Leonhardt. Remote monitoring of breathing dynamics using infrared thermography. *Biomedical Optics Express*, 6(11): 4378–4394, October 2015. ISSN 2156-7085. doi: 10.1364/BOE.6.004378. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4646547/.

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

[46] Dvijesh Shastri, Manos Papadakis, Panagiotis Tsiamyrtzis, Barbara Bass, and Ioannis Pavlidis. Perinasal Imaging of Physiological Stress and Its Affective Potential. *IEEE Transactions on Affective Computing*, 3(3): 366–378, July 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2012.13. Conference Name: IEEE Transactions on Affective Computing.

[47] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0197-0. URL https://doi.org/10.1186/s40537-019-0197-0.

[48] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning From Simulated and Unsupervised Images Through Adversarial Training. In *IEEE Conference on CVPR*, pages 2107–2116, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Shrivastava_Learning_From_Simulated_CVPR_2017_paper.html.

[49] Nahian Siddique, Paheding Sidike, Colin Elkin, and Vijay Devabhaktuni. U-Net and its variants for medical image segmentation: theory and applications. *IEEE Access*, 9:82031–82057, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3086020. URL http://arxiv.org/abs/2011.01118. ZSCC: 0000002 arXiv: 2011.01118.

[50] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science, pages 240–248, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67558-9. doi: 10.1007/978-3-319-67558-9_28. ZSCC: NoCitationData[s0].

[51] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-Resolution Representations for Labeling Pixels and Regions. *arXiv:1904.04514 [cs]*, April 2019. URL http://arxiv.org/abs/1904.04514. ZSCC: 0000335 arXiv: 1904.04514.

[52] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring Cross-Image Pixel Contrast for Semantic Segmentation. pages 7303–7313, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Wang_Exploring_Cross-Image_Pixel_Contrast_for_Semantic_Segmentation_ICCV_2021_paper.html.

[53] W. K. Wong and Haitao Zhao. Eyeglasses removal of thermal image based on visible information. *Information Fusion*, 14(2):163–176, April 2013. ISSN 1566-2535. doi: 10.1016/j.inffus.2011.09.002. URL https://doi.org/10.1016/j.inffus.2011.09.002. ZSCC: 0000026.

[54] Yuan Xue, Tao Xu, and Xiaolei Huang. Adversarial learning with multi-scale loss for skin lesion segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 859–863, April 2018. doi: 10.1109/ISBI.2018.8363707. ISSN: 1945-8452.

[55] Yuan Xue, Tao Xu, Han Zhang, L. Rodney Long, and Xiaolei Huang. SegAN: Adversarial Network with Multi-scale L1 Loss for Medical Image Segmentation. *Neuroinformatics*, 16(3):383–392, October 2018. ISSN 1559-0089. doi: 10.1007/s12021-018-9377-x. URL https://doi.org/10.1007/s12021-018-9377-x.

[56] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang. Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation, April 2021. URL http://arxiv.org/abs/1909.11065. arXiv:1909.11065 [cs].

[57] Chaoyi Zhang, Yang Song, Sidong Liu, Scott Lill, Chenyu Wang, Zihao Tang, Yuyi You, Yang Gao, Alexander Klistorner, Michael Barnett, and Weidong Cai. MS-GAN: GAN-Based Semantic Segmentation of Multiple Sclerosis Lesions in Brain Magnetic Resonance Imaging. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, December 2018. doi: 10.1109/DICTA.2018.8615771. ZSCC: 0000016.

[58] Feihu Zhang, Philip Torr, Rene Ranftl, and Stephan R. Richter. Looking Beyond Single Images for Contrastive Semantic Segmentation Learning. May 2021. URL https://openreview.net/forum?id=MSVlSMBbBt. ZSCC: 0000000.

[59] Xinming Zhang, Xiaobin Zhu, 3rd Xiao–Yu Zhang, Naiguang Zhang, Peng Li, and Lei Wang. SegGAN: Semantic Segmentation with Generative Adversarial Network. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–5, September 2018. doi: 10.1109/BigMM.2018.8499105. ZSCC: 0000018.

[60] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region Mutual Information Loss for Semantic Segmentation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a67c8c9a961b4182688768dd9ba015fe-Abstract.html.

[61] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive Learning for Label Efficient Semantic Segmentation. pages 10623–10633, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhao_Contrastive_Learning_for_Label_Efficient_Semantic_Segmentation_ICCV_2021_paper.html.

[62] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, June 2020. ISSN 1558-254X. doi: 10.1109/TMI.2019.2959609. Conference Name: IEEE Transactions on Medical Imaging.

[63] Juntang Zhuang. LadderNet: Multi-path networks based on U-Net for medical image segmentation. *arXiv:1810.07810 [cs, eess]*, August 2019. URL http://arxiv.org/abs/1810.07810. ZSCC: NoCitationData[s0] arXiv: 1810.07810.

[64] Long Zhuo, Shunquan Tan, Bin Li, and Jiwu Huang. Self-Adversarial Training Incorporating Forgery Attention for Image Forgery Localization. *IEEE Transactions on Information Forensics and Security*, 17:819–834, 2022. ISSN 1556-6021. doi: 10.1109/TIFS.2022.3152362.