2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506219

# TWO-PHASE FEATURE FUSION NETWORK FOR VISIBLE-INFRARED PERSON RE-IDENTIFICATION

*Yunzhou Cheng*[1], *Guoqiang Xiao*[1⋆], *Xiaoqin Tang*[1], *Wenzhuo Ma*[2], *Xinye Gou*[2]

[1]College of Computer and Information Science, Southwest University, Chongqing 400715, China
[2]Chongqing Productivity Council, Chongqing 401120, China
⋆ *Corresponding author: gqxiao@swu.edu.cn*

## ABSTRACT

Visible-infrared person re-identification (VI-ReID) is a challenging problem that aims to match pedestrians captured by visible and infrared cameras. Prevailing methods in this field mainly focus on learning sharable feature representations from the last layer of deep convolution neural networks(CNNs). However, due to the large intra-modality variations and cross-modality variations, the last layer's sharable feature representations are less discriminative. To remedy this, we propose a novel Two-Phase Feature Fusion Network(TFFN) to enhance the discriminative feature learning via feature fusion. Specifically, TFFN contains two fusion modules: (1) Multi-Level Fusion Module(MLFM) that reweights and fuses intra-modality multi-level features to utilize high- and low-level information; (2) Graph-Level Fusion Module (GLFM) that mines and fuses rich mutual information across the two modalities by employing a cross-modality graph attention network. Additionally, for effective fusion, we develop a deep supervision method to enhance the discrimination of pre-fusion features and eliminate noise information. Extensive experiments show that TFFN outperforms the state-of-the-art methods on two mainstream VI-ReID datasets: SYSU-MM01 and RegDB.

***Index Terms***— Person re-identification, cross-modality, feature fusion, deep supervision

## 1. INTRODUCTION

Person re-identification(ReID) is a pedestrian retrieval problem. Powered by deep learning technology in recent years, some methods [1, 2] have achieved human-level performance. However, most of these methods only consider images of people collected by visible (RGB) cameras during the day-time. Thus, they can't handle some night-time tasks where images are collected by infrared (IR) cameras under poor illumination conditions.

To alleviate this problem, cross-modality visible-infrared person re-identification (VI-ReID) arises naturally. Given a query RGB (or IR) image, the goal is to match the corresponding person in a gallery of IR (or RGB) images. However, VI-ReID faces far greater challenges than visible-visible ReID. As shown in Fig. 1, VI-ReID not only encounters the intra-modality variations caused by different camera views, but also suffers from the large cross-modality variations resulting from different camera spectrums (visible and infrared).

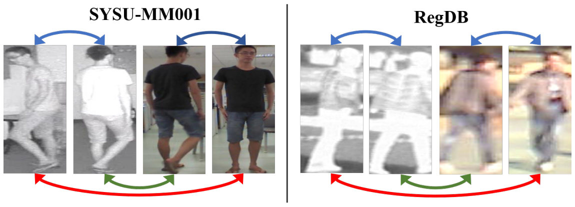

**Fig. 1**. Challenges in VI-ReID. Blue arrow: Intra-modality variations. Green arrow: Cross-modality variations. Red arrow: Both. Images are taken from the SYSU-MM01 [3] and RegDB [4] datasets.

Therefore, how to learn discriminative feature representations to deal with the aforementioned two types of variations simultaneously becomes a thorny problem. Existing methods in this field mainly focus on learning high-level sharable feature representations of the two modalities, aiming to learn invariant shared features while preserving high discriminability, including one-stream [3, 5] networks, as well as the recently popular two- [6, 7] or three-stream [8] networks. Besides, some scholars use Generative Adversarial Networks(GANs) [7, 9] or investigate a better loss function [10] to deal with the modality variations.

However, there are two drawbacks for these methods, leading to a lack of discrimination of the learned features. One drawback is that the previous researches focus more on high-level semantic feature learning, rather than low-level details. This, in some sense, might not be a good fit for VI-ReID. Because with the color loss for IR images, low-level details can be a good complement for the feature representations. In addition, most existing studies conduct research on feature-level sharable representations across the two modalities, yet fewer research on mutual representations between images of the same identity. This rich mutual information

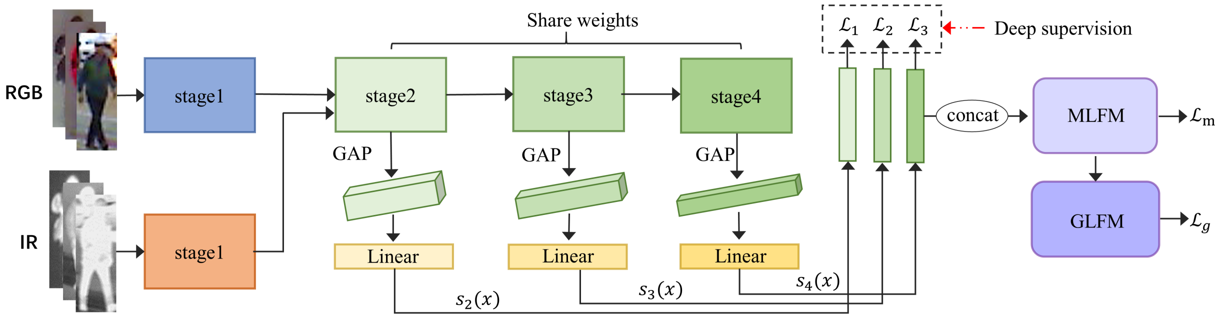

**Fig. 2**. The proposed two-phase feature fusion network (TFFN) for VI-ReID.

reduces the modality variations at the training process to enhance the discriminative feature learning.

To overcome the two drawbacks, we propose a novel two-phase feature fusion network(TFFN) to enhance the discrimination of the learned features by using feature fusion. With its initial application for object detection and image segmentation [11, 12], feature fusion is first used for VI-ReID in this paper. As shown in Fig. 2, the backbone network with deep supervision is used for discriminative feature embedding, and the two feature fusion modules, MLFM and GLFM, are placed at the end of the backbone network. MLFM is designed to utilize high- and low-level information, and GLFM fuses the output of MLFM to exploit graph-level rich mutual information.

The main contribution of this paper is summarized as follows: (1) The TFFN, with deep supervision, is proposed to enhance the discrimination level of features for VI-ReID. (2) In order to acquire the intra-modality multi-level information, an IMMF is utilized, whilst an MLFM is designed to exploit the graph-level rich mutual information across the two modalities. (3) The experimental results indicate that the presented methodology outperforms the compared state-of-the-art methods on two mainstream datasets.

## 2. PROPOSED METHOD

In this section, we first introduce the backbone network with deep supervision for feature embedding, and then describe MLFM and GLFM in detail. Finally, we introduce the overall loss function of TFFN.

### 2.1. Deep Supervision for Feature Embedding

The backbone network is ResNet50 [13] that basically contains four convolutional stages. In order to learn the shared feature embedding, we set the parameters of the last three stages to be shared for the two modalities.

At each training batch, we use an identity-balanced sampling strategy [10], first randomly selecting $N$ identities, then selecting $NM$ visible images of $N$ identities as input to the RGB branch, and $NM$ infrared images as input to the IR branch ($N = M = 4$ in our experiments). For each image, we take the output from each modality-shared stage and use global average pooling(GAP) to produce an embedding at each stage. At the same time, each branch contains a fully connected layer, which bring all embeddings to the same dimension. For clarification, we denote the embedding that is produced at stage $i$ as $s_i(x)$, where $C$ is the embedding dimension ($C = 1024$ in our experiments).

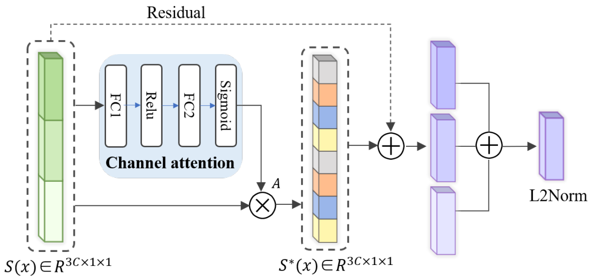
Residual
FC1
Relu
FC2
Sigmoid
Channel attention
$A$
$S(x) \in R^{3C\times1\times1}$
$S^{*}(x) \in R^{3C\times1\times1}$
L2Norm

**Fig. 3**. The structure of MLFM. $\otimes$ and $\oplus$ indicate the element-wise multiplication and addition, respectively.

Inspired by [14], we add an auxiliary supervision branch on top of each modality-shared stage as deep supervision, rather than simply adding supervision to the last layer of the network. Additionally, we optimize each deep supervision branch with a combination of an identity loss and a hard-mining triplet loss [15]. It is formulated by:

$$\mathcal{L}_i = \mathcal{L}_{id} + \mathcal{L}_{tri}. \quad (1)$$

The identity loss $\mathcal{L}_{id}$ is cross entropy loss, which enables the network to learn the ability to classify, and the triplet loss $\mathcal{L}_{tri}$ narrows the intra-class distance while increases the inter-class distance. The pre-fusion features, in this way, are more discriminative, and the transmission error between layers is further minimized.

### 2.2. Multi-level Feature Fusion

MLFM, illustrated in Fig. 3, is placed at the end of the backbone network to reweight and fuse multi-level features, taking

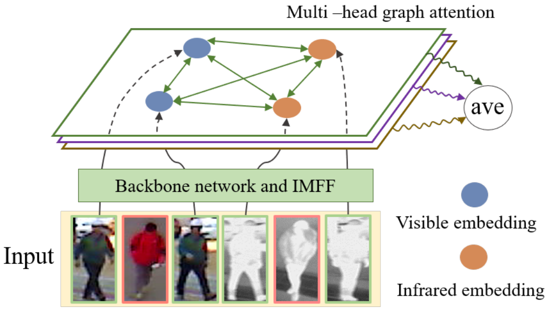

**Fig. 4**. The structure of GLFM.

high- and low- level information into account.

We firstly concatenate the embeddings produced at each modality-shared stage, obtaining a multi-level feature $S(x) \in \mathbb{R}^{3C\times 1}$. Afterwards, a global attention mask is dynamically calculated to model the inter-dependencies between the channel of $S(x)$. It improves the performance of learning discriminative features by assigning greater weights to important channels, similar to [16]. Meanwhile, we use a residual connection of the original multi-level feature $S(x)$ to save the primary multi-level feature and stabilize the training process. In summary, for an input $S(x)$, the boosted feature $S^*(x)$ is computed as:

$$A = Sigmoid(W_2(ReLU(W_1(S(x))))), \quad (2)$$

$$S^*(x) = A \otimes S(x) \oplus S(x), \quad (3)$$

where $W_1$ and $W_2$ are the parameters of two fully connected layers (FC), respectively.

Finally, the output of MLFM is the element-wise addition of the boosted features $\{s^*(x)_i\}_{i=2}^{4}$from each modality-shared stage. Besides, the obtained feature is $l_2$ normalized for stable convergence.

### 2.3. Graph-level Feature Fusion

The input of GLFM is the output of MLFM, with each input uniquely representing the feature of an individual image. As shown in Fig. 4, we regard the input features as a set of nodes, where only nodes with the same ID have edge connections. GLFM fuses all node features that belong to the same identity to integrate graph-level rich mutual information across the two modalities, which reduces the modality variations to enhance the discriminative feature learning.

Given the input node features, $\left\{f_k \in \mathbb{R}^{C\times 1}\right\}_{k=1}^{K}$, where $K$ is the batch size, and each node feature corresponds to a ground-truth label $y_k$. We adopt the graph attention network [17] to reveal the mutual relations between these node features. Thus, we get the attention coefficient $s_{ij}$ using:

$$s_{ij} = \frac{\exp(\sigma([h(f_i)\|h(f_j)] \cdot \mathbf{w})}{\sum_{\forall y_i=y_k} \exp(\sigma([h(f_i)\|h(f_k)] \cdot \mathbf{w})}, \quad (4)$$

where $s_{ij}$ measures the $j^{th}$ node feature impact on $i^{th}$ node feature. $\sigma(\cdot)$ is the LeakyRelu function. $\|$ represents the concatenation operation. $h(\cdot)$ is a $1 \times 1$ convolutional layer to convert the node features to a lower-dimensional feature space from $C$ to $C/2$. $\mathbf{w} \in \mathbb{R}^{C\times 1}$ is a learnable parameter vector.

We also employ multi-head attention [18] to stabilize the training process, which performs $M$ independent graph attention networks ($M = 4$ in our experiments), and then perform an average operation on the obtained features:

$$f_i = \frac{1}{M}\sum_{m=1}^{M}\sum\nolimits_{\forall y_i=y_j} s_{ij}^{m} \cdot f_j. \quad (5)$$

### 2.4. Loss Function

In order to further improve the discrimination of the learned features, we optimize each fusion module with a combination of an identity loss and a triplet loss, which is the same as the loss function $\mathcal{L}_i$ for each deep supervision branch.

Therefore, the overall loss function consists of three parts:

$$\mathcal{L}_{all} = \sum_{i=1}^{3}\mathcal{L}_i + \mathcal{L}_m + \lambda\mathcal{L}_g, \quad (6)$$

where $\mathcal{L}_m$ and $\mathcal{L}_g$ are the loss functions for MLFM and GLFM, respectively. $\lambda$ is an adaptive coefficient to stabilize GLFM at the early stage of the training process, which is initialized as 0 and increases gradually [19].

## 3. EXPERIMENTS

### 3.1. Datasets and Settings

SYSU-MM01 [3] and RegDB [4] are used for our experiments. **SYSU-MM01** is captured by six cameras (four RGB and two IR). Referring to the SYSU-MM01 evaluation criterion [3], the training set consists of 22,258 RGB images and 11909 IR images of 395 identities. For testing, we randomly chose 3803 IR images of 96 identities as the queries and 301 RGB images as the gallery. **RegDB** is collected by a dual-camera system, which contains one RGB camera and one IR camera. With the RegDB evaluation protocol [20], 2060 RGB images and 2060 IR images of 206 identities compose the training set, whist the same number of images and identities as query and gallery set for testing. In the end, the rank-k matching accuracy and mean Average Precision (mAP) are employed to evaluate the performance.

We use Pytorch to implement the proposed method. All the images are resized to 288×144×3. The data augmentation includes random cropping, horizontal flipping, and random erasing [21]. The backbone network is pre-trained on the ImageNet dataset. We select SGD optimizer with the momentum of 0.9 as the optimizer. The model is trained for 70 epochs. The initial learning rate is set to 0.01, which is adjusted by a

**Table 1**. Comparision on RegDB dataset.

| Methods | $r$=1 | $r$=10 | mAP | refrence |
|---|---|---|---|---|
| eBDTR [10] | 34.62 | 58.96 | 33.46 | TIFS2019 |
| MSR [6] | 48.43 | 70.32 | 48.67 | TIP2019 |
| DSCSN [7] | 60.8 | 78.6 | 60.0 | AAAI2020 |
| Xmodal [24] | 62.21 | 83.13 | 60.18 | AAAI2020 |
| DDAG [19] | 69.34 | 86.19 | 63.46 | ECCV2020 |
| Hi-CMD [23] | 70.93 | 86.39 | 66.04 | CVPR2020 |
| **TFFN** | **81.17** | **93.69** | **77.16** | This paper |

**Table 2**. Comparision on SYSU-MM01 dataset.

| Methods | $r$=1 | $r$=10 | mAP | refrence |
|---|---|---|---|---|
| eBDTR [10] | 27.82 | 67.34 | 28.42 | TIFS2019 |
| Hi-CMD [23] | 35.1 | 77.6 | 37.4 | CVPR2020 |
| AlignGAN [9] | 42.4 | 85.0 | 40.7 | ICCV2019 |
| AGW [1] | 47.50 | 84.39 | 47.65 | arXiv2020 |
| Xmodal [24] | 49.92 | 89.79 | 50.73 | AAAI2020 |
| DDAG [19] | 54.75 | 90.39 | 53.20 | ECCV2020 |
| **TFFN** | **58.37** | **91.30** | **56.02** | This paper |

**Table 3**. Ablation study on SYSU-MM01 dataset.

| Methods | $r$=1 | $r$=10 | mAP |
|---|---|---|---|
| B | 49.41 | 87.67 | 48.48 |
| B(deep) | 52.56 | 88.11 | 51.32 |
| B(deep)+MLFM | 56.06 | 90.30 | 53.76 |
| B(deep)+GLFM | 54.46 | 88.38 | 52.15 |
| B(deep)+MLFM+GLFM | 58.37 | 91.30 | 56.02 |
| B(deep)+MLFM w/o A | 55.38 | 89.51 | 52.98 |
| B(deep)+GLFM w/o H | 53.51 | 88.27 | 51.89 |

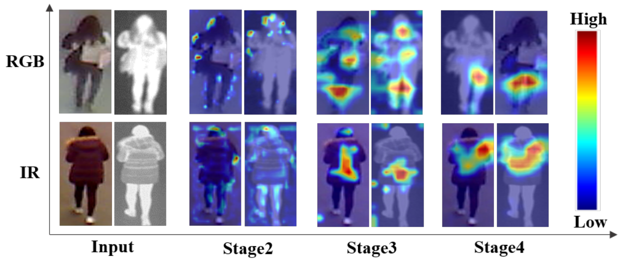

**Fig. 5**. The results of Grad-CAM [25] visualization.

warm-up strategy [22], and decays by 0.1 and 0.01 in the 30th and 50th epoch, respectively.

### 3.2. Comparison with State-of-the-art Methods

In this subsection, the proposed TFFN and the state-of-the-art VI-ReID methods are compared. TFFN achieves the improvement of 10.24% Rank-1 and 11.12% mAP above Hi-CMD [23] on RegDB dataset in Table 1 and 3.62% Rank-1 and 2.82% mAP above DDAG [19] on SYSU-MM01 dataset in Table 2. Although the Hi-CMD and DDAG methods both consider the feature- and image-level constraints, the low-level detail information at the feature-level and the rich mutual information at the image-level were not properly taken into account. Differently, in this work, we fuse both information to discriminate features that are key to VI-ReID, so far achieving the best performance in comparison to some of the existing methods.

### 3.3. Ablation and Visualization

**Ablation.** we conduct a detailed ablation study to demonstrate the effectiveness of each component on the SYSU-MM01 dataset. "B" denotes our baseline network with supervision on the last layer. "w/o A"("w/o H") indicates that our model is trained without global channel attention in MLFM (multi-head attention in GLFM).

From the results in Table 3, we can see that the deep supervision method improves the performance compared with the baseline. On this basis, MLFM and GLFM further improve the performance, respectively. By combining the two fusion modules, we achieve the best performance. Subsequently, we remove the global channel attention in MLFM and the multi-head attention in GLFM, both of which result in performance reduction. This suggests that both types of attention modules promote the efficiency of feature fusion.

**Visualization.** We adopt Grad-CAM [25] method to locate the important regions of each modality-shared stage, as depicted in Fig. 5. We can find that different stages truly focus on different locations. By fusing the multi-level features, we can enhance the features with more discriminability.

## 4. CONCLUSION

We present a two-phase feature fusion network(TFFN) for VI-ReID. TFFN is innovative in two aspects: its MLFM utilizes high- and low-level information within each modality to directly enhance the feature representation; the GLFM incorporates image-level mutual information across the two modalities to enhance the discriminative feature learning. We further develop a deep supervision method to enhance the discrimination of pre-fusion features and eliminate noise information. The experiment results demonstrate that our method achieves outstanding performance. We hope the proposed TFFN can be a useful baseline for future researches on feature fusion for VI-ReID.

## 5. ACKNOWLEDGEMENT

This work is supported by the Natural Science Foundation of Chongqing, China (Grant No. cstc2019jcyj-zdxmX0025).

## 6. REFERENCES

[1] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020.

[2] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[3] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai, "Rgb-infrared cross-modality person re-identification," in *ICCV*, 2017.

[4] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, 2017.

[5] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *CVPR*, 2019.

[6] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Transactions on Image Processing*, vol. 29, 2019.

[7] Shizhou Zhang, Yifei Yang, Peng Wang, Xiuwei Zhang, and Yanning Zhang, "Attend to the difference: Cross-modality person re-identification via contrastive correlation," *arXiv preprint arXiv:1910.11656*, 2019.

[8] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu, "Cross-modality person re-identification with shared-specific feature transfer," in *CVPR*, 2020.

[9] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *ICCV*, 2019.

[10] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 15, 2019.

[11] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Hypercolumns for object segmentation and fine-grained localization," in *CVPR*, 2015.

[12] Saining Xie and Zhuowen Tu, "Holistically-nested edge detection," in *CVPR*, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[14] Dawei Sun, Anbang Yao, Aojun Zhou, and Hao Zhao, "Deeply-supervised knowledge synergy," in *CVPR*, 2019.

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[16] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[17] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[18] Guocong Song and Wei Chai, "Collaborative learning for deep neural networks," in *NIPS*, 2018, pp. 1832–1841.

[19] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," *arXiv preprint arXiv:2007.09314*, 2020.

[20] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *AAAI*, 2018.

[21] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang, "Random erasing data augmentation.," in *AAAI*, 2020.

[22] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, 2019.

[23] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim, "Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *CVPR*, 2020.

[24] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong, "Infrared-visible cross-modal person re-identification with an x modality.," in *AAAI*, 2020.

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.